

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013 Transcript – Recitation: Geniuses and Chocolates

Hi. Today, we're going to do a really fun problem called geniuses and chocolates. And what this problem is exercising is your knowledge of properties of probability laws. So let me just clarify what I mean by that. Hopefully, by this point, you have already learned what the axioms of probability are. And properties of probability laws are essentially any rules that you can derive from those axioms.

So take for example the fact that the probability of $A \cup B$ is equal to the probability of A plus the probability of B minus the probability of the intersection. That's an example of a property of a probability law. So enough with the preamble. Let's see what the problem is asking us. In this problem, we have a class of students. And we're told that 60% of the students are geniuses. 70% of the students love chocolate. So I would be in that category. And 40% fall into both categories.

And our job is to determine the probability that a randomly selected student is neither a genius nor a chocolate lover. So first I just want to write down the information that we're given in the problem statement. So if you let G denote the event that a randomly selected student is a genius then the problem statement tells us that the probability of G is equal to 0.6.

Similarly, if we let C denote the event that a randomly selected student is a chocolate lover, then we have that the probability of C is equal to 0.7. Lastly, we are told that the probability a randomly selected student falls into both categories is 0.4. And the way we can express that using the notation already on the board is probability of $G \cap C$ is equal to 0.4.

OK, now one way of approaching this problem is to essentially use this information and sort of massage it using properties of probability laws to get to our answer. Instead, I'm going to take a different approach, which I think will be helpful. So namely, we're going to use something called a Venn diagram. Now a Venn diagram is just a tool that's really useful for telling you how different sets relate to each other and how their corresponding probabilities relate to each other.

So the way you usually draw this is you draw a rectangle, which denotes your sample space, which of course, we call Ω . And then you draw two intersecting circles. So one to represent our geniuses and one to represent our chocolate lovers. And the reason why I drew them intersecting is because we know that there are 40% of the students in our class are both geniuses and chocolate lovers. OK, and the way you sort of interpret this diagram is the space outside these two circles correspond to students who are neither geniuses nor chocolate lovers. And so just keep in mind that the probability corresponding to these students on the outside, that's actually what we're looking for.

Similarly, students in this little shape, this tear drop in the middle, those would correspond to geniuses and chocolate lovers. You probably get the idea. So this is our Venn diagram. Now I'm going to give you guys a second trick if you will. And that is to work with partitions.

So I believe you've seen partitions in lecture by now. And a partition is essentially a way of cutting up the sample space into pieces. But you need two properties to be true. So the pieces that you cut up your sample space into, they need to be disjoint, so they can't overlap. So for instance, G and C are not disjoint because they overlap in this tear drop region.

Now the second thing that a partition has to satisfy is that if you put all the pieces together, they have to comprise the entire sample space. So I'm just going to put these labels down on my graph. X, Y, Z, and W. So X is everything outside the two circles but inside the rectangle. And just note, again, that what we're actually trying to solve in this problem is the probability of X, the probability that you're neither genius, because you're not in this circle, and you're not a chocolate lover, because you're not in this circle.

So Y I'm using to refer to this sort of crescent moon shape. Z, I'm using to refer to this tear drop. And W, I'm using to refer to this shape. So, hopefully, you agree that X, Y, Z, and W form a partition because they don't overlap. So they are disjoint. And together they form omega.

So now we're ready to do some computation. The first step is to sort of get the information we have written down here in terms of these new labels. So hopefully, you guys buy that G is just the union of Y and Z. And because Y and Z are disjoint, we get that the probability of the union is the sum of the probabilities. And, of course, we have from before that this is 0.6.

Similarly, we have that the probability of C is equal to the probability of Z union W. And, again, using the fact that these two guys are disjoint, you get this expression. And that is equal to 0.7.

OK, and the last piece of information, G intersects C corresponds to Z, or our tear drop, and so we have that the probability of Z is equal to 0.4. And now, if you notice, probability of Z shows up in these two equations. So we can just plug it in. So plug in 0.4 into this equation. We get P of Y plus 0.4 is 0.6. So that implies that P of Y is 0.2. That's just algebra. And similarly we have point. 0.4 plus P of W is equal to 0.7. So that implies that P of W is 0.3. Again, that's just algebra.

So now we're doing really well because we have a lot of information. We know the probability of Y, the probability of Z, the probability of W. But remember we're going for, we're trying to find the probability of X. So the way we finally put all this information together to solve for X is we use the axiom that tells us that 1 is equal to the probability of the sample space. And then, again, we're going to use sort of this really helpful fact that X, Y, Z, and W form a partition of omega to go ahead and write this as probability of X plus probability of Y plus probability, oops, I made a mistake. Hopefully, you guys caught that. It's really, oh, no. I'm right. Never mind.

Probability of X plus probability of Y plus probability of Z plus probability of W. And now we can go ahead and plug-in the values that we solved for previously. So we get probability of X plus 0.2 plus 0.4 plus 0.3. These guys sum to 0.9. So, again, just simple arithmetic, we get that the probability of X is equal to 0.1.

So we're done because we've successfully found that the probability that a randomly selected student is neither a genius nor a chocolate lover is 0.1. So this was a fairly straightforward

problem. But there are some important takeaways. The first one is that Venn diagrams are a really nice tool. Whenever the problem is asking you how different sets relate to each other or how different probabilities relate to each other, you should probably draw Venn diagram because it will help you.

And the second takeaway is that it's frequently useful to divide your sample space into a partition mainly because sort of the pieces that compose a partition are disjoint. So we will be back soon to solve more problems.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013 Transcript – Recitation: The Probability of the Difference of Two Events

Hi. In this problem, we're going to use the set of probability axioms to derive the probability of the difference of two events. Now, before we get started, there's one thing you might notice that, the equation we're trying to prove is actually quite complicated. And I don't like it either, so the first thing we're going to do will be to find a simpler notation for the events that we're interested in.

So we start with two events, A and B, and there might be some intersection between the two events. We'll label the set of points or samples in A that are not in B, as a set C. So C will be A intersection B complement. Similarly, for all points that are in B but not in A, this area, we'll call it D.

And D will be the set A complement intersection B. And finally, for points that are in the intersection of A and B, we'll call it E. So E is A intersection B. And for the rest of our problem, we're going to be using the notation C, D, and E instead of whatever's down below.

If we use this notation, we can rewrite our objective as the following. We want to show that the probability of C union D is equal to the probability of the event A plus the probability of B minus twice the probability of E. And that will be our goal for the problem.

Now, let's take a minute to review what the axioms are, what the probability axioms are. The first one says non-negativity. We take any event A, then the probability of A must be at least 0. The second normalization says the probability of the entire space, the entire sample space omega, must be equal to 1.

And finally, the additivity axiom, which will be the axiom that we're going to use for this problem says, if there are two events, A and B that are disjoint-- which means they don't have anything in common, therefore, the intersection is the empty set. Then the probability of their union will be equal to the probably A plus the probability of B. For the rest of the problem, I will refer to this axiom as add.

So whenever we invoke this axiom, I'll write "add" on the board. Let's get started. First, we'll invoke the additivity axioms to argue that the probability of C union D is simply the sum of probability of C plus probability of D. Why is this true?

We can apply this axiom, because the set C here and the set D here, they're completely disjoint from each other. And in a similar way, we'll also notice the following. We see that A is equal to the union of the set C and E.

And also, C and E, they're disjoint with each other, because C and E by definition don't share any points. And therefore, we have probably A is equal to probability of C plus the probability of E.

Now, in a similar way, the probability of event B can also be written as a probability of D plus the probability of E, because event B is the union of D and E.

And D and E are disjoint from each other. So we again invoke the additivity axiom. Now, this should be enough to prove our final claim. We have the probability of C union D. By the very first line, we see this is simply probability of C plus the probability of D.

Now, I'm going to insert two terms here to make the connection with a second part of the equation more obvious. That is, I will write probability C plus probability E plus probability D plus probability of E. Now, I've just added two terms here-- probability E. So to make the equality valid or subtract it out two times, the probability of E.

Hence this equality is valid. So if we look at this equation, we see that there are two parts here that we've already seen before right here. The very first parenthesis is equal to the probability of A. And the value of the second parenthesis is equal to the probability of B. We just derived these here.

And finally, we have the minus 2 probability of E. This line plus this line gives us the final equation. And that will be the answer for the problem.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013

Transcript – Lecture 1

The following content is provided under a Creative Commons license. Your support will help MIT OpenCourseWare continue to offer high quality educational resources for free. To make a donation or view additional materials from hundreds of MIT courses, visit MIT OpenCourseWare at ocw.mit.edu.

PROFESSOR: OK, so welcome to 6.041/6.431, the class on probability models and the like. I'm John Tsitsiklis. I will be teaching this class, and I'm looking forward to this being an enjoyable and also useful experience. We have a fair amount of staff involved in this course, your recitation instructors and also a bunch of TAs, but I want to single out our head TA, Uzoma, who is the key person in this class. Everything has to go through him. If he doesn't know in which recitation section you are, then simply you do not exist, so keep that in mind.

All right. So we want to jump right into the subject, but I'm going to take just a few minutes to talk about a few administrative details and how the course is run. So we're going to have lectures twice a week and I'm going to use old fashioned transparencies. Now, you get copies of these slides with plenty of space for you to keep notes on them. A useful way of making good use of the slides is to use them as a sort of mnemonic summary of what happens in lecture. Not everything that I'm going to say is, of course, on the slides, but by looking them you get the sense of what's happening right now. And it may be a good idea to review them before you go to recitation.

So what happens in recitation? In recitation, your recitation instructor is going to maybe review some of the theory and then solve some problems for you. And then you have tutorials where you meet in very small groups together with your TA. And what happens in tutorials is that you actually do the problem solving with the help of your TA and the help of your classmates in your tutorial section.

Now probability is a tricky subject. You may be reading the text, listening to lectures, everything makes perfect sense, and so on, but until you actually sit down and try to solve problems, you don't quite appreciate the subtleties and the difficulties that are involved. So problem solving is a key part of this class. And tutorials are extremely useful just for this reason because that's where you actually get the practice of solving problems on your own, as opposed to seeing someone else who's solving them for you.

OK but, mechanics, a key part of what's going to happen today is that you will turn in your schedule forms that are at the end of the handout that you have in your hands. Then, the TAs will be working frantically through the night, and they're going to be producing a list of who goes into what section. And when that happens, any person in this class, with probability 90%, is going to be happy with their assignment and, with probability 10%, they're going to be unhappy.

Now, unhappy people have an option, though. You can resubmit your form together with your full schedule and constraints, give it back to the head TA, who will then do some further juggling and reassign people, and after that happens, 90% of those unhappy people will become happy. And 10% of them will be less unhappy.

OK. So what's the probability that a random person is going to be unhappy at the end of this process? It's 1%. Excellent. Good. Maybe you don't need this class. OK, so 1%. We have about 100 people in this class, so there's going to be about one unhappy person. I mean, anywhere you look in life, in any group you look at, there's always one unhappy person, right? So, what can we do about it?

All right. Another important part about mechanics is to read carefully the statement that we have about collaboration, academic honesty, and all that. You're encouraged, it's a very good idea to work with other students. You can consult sources that are out there, but when you sit down and write your solutions you have to do that by setting things aside and just write them on your own. You cannot copy something that somebody else has given to you.

One reason is that we're not going to like it when it happens, and then another reason is that you're not going to do yourself any favor. Really the only way to do well in this class is to get a lot of practice by solving problems yourselves. So if you don't do that on your own, then when quiz and exam time comes, things are going to be difficult.

So, as I mentioned here, we're going to have recitation sections, that some of them are for 6.041 students, some are for 6.431 students, the graduate section of the class. Now undergraduates can sit in the graduate recitation sections. What's going to happen there is that things may be just a little faster and you may be covering a problem that's a little more advanced and is not covered in the undergrad sections. But if you sit in the graduate section, and you're an undergraduate, you're still just responsible for the undergraduate material. That is, you can just do the undergraduate work in the class, but maybe be exposed at the different section.

OK. A few words about the style of this class. We want to focus on basic ideas and concepts. There's going to be lots of formulas, but what we try to do in this class is to actually have you understand what those formulas mean. And, in a year from now when almost all of the formulas have been wiped out from your memory, you still have the basic concepts. You can understand them, so when you look things up again, they will still make sense. It's not the plug and chug kind of class where you're given a list of formulas, you're given numbers, and you plug in and you get answers.

The really hard part is usually to choose which formulas you're going to use. You need judgment, you need intuition. Lots of probability problems, at least the interesting ones, often have lots of different solutions. Some are extremely long, some are extremely short. The extremely short ones usually involve some kind of deeper understanding of what's going on so that you can pick a shortcut and use it. And hopefully you are going to develop this skill during this class.

Now, I could spend a lot of time in this lecture talking about why the subject is important. I'll keep it short because I think it's almost obvious. Anything that happens in life is uncertain. There's uncertainty anywhere, so whatever you try to do, you need to have some way of dealing or thinking about this uncertainty. And the way to do that in a systematic way is by using the models that are given to us by probability theory.

So if you're an engineer and you're dealing with a communication system or signal processing, basically you're facing a fight against noise. Noise is random, is uncertain. How do you model it? How do you deal with it?

If you're a manager, I guess you're dealing with customer demand, which is, of course, random. Or you're dealing with the stock market, which is definitely random. Or you play the casino, which is, again, random, and so on. And the same goes for pretty much any other field that you can think of.

But, independent of which field you're coming from, the basic concepts and tools are really all the same. So you may see in bookstores that there are books, probability for scientists, probability for engineers, probability for social scientists, probability for astrologists. Well, what all those books have inside them is exactly the same models, the same equations, the same problems. They just make them somewhat different word problems.

The basic concepts are just one and the same, and we'll take this as an excuse for not going too much into specific domain applications. We will have problems and examples that are motivated, in some loose sense, from real world situations. But we're not really trying in this class to develop the skills for domain-specific problems. Rather, we're going to try to stick to general understanding of the subject.

OK. So the next slide, of which you do have in your handout, gives you a few more details about the class. Maybe one thing to comment here is that you do need to read the text. And with calculus books, perhaps you can live with a just a two page summary of all of the interesting formulas in calculus, and you can get by just with those formulas. But here, because we want to develop concepts and intuition, actually reading words, as opposed to just browsing through equations, does make a difference.

In the beginning, the class is kind of easy. When we deal with discrete probability, that's the material until our first quiz, and some of you may get by without being too systematic about following the material. But it does get substantially harder afterwards. And I would keep restating that you do have to read the text to really understand the material.

OK. So now we can start with the real part of the lecture. Let us set the goals for today. So probability, or probability theory, is a framework for dealing with uncertainty, for dealing with situations in which we have some kind of randomness. So what we want to do is, by the end of today's lecture, to give you anything that you need to know how to set up what does it take to set up a probabilistic model. And what are the basic rules of the game for dealing with probabilistic models? So, by the end of this lecture, you will have essentially recovered half of this semester's tuition, right?

So we're going to talk about probabilistic models in more detail-- the sample space, which is basically a description of all the things that may happen during a random experiment, and the probability law, which describes our beliefs about which outcomes are more likely to occur compared to other outcomes. Probability laws have to obey certain properties that we call the axioms of probability. So the main part of today's lecture is to describe those axioms, which are the rules of the game, and consider a few really trivial examples.

OK, so let's start with our agenda. The first piece in a probabilistic model is a description of the sample space of an experiment. So we do an experiment, and by experiment we just mean that just something happens out there. And that something that happens, it could be flipping a coin, or it could be rolling a dice, or it could be doing something in a card game.

So we fix a particular experiment. And we come up with a list of all the possible things that may happen during this experiment. So we write down a list of all the possible outcomes. So here's a list of all the possible outcomes of the experiment. I use the word "list," but, if you want to be a little more formal, it's better to think of that list as a set.

So we have a set. That set is our sample space. And it's a set whose elements are the possible outcomes of the experiment. So, for example, if you're dealing with flipping a coin, your sample space would be heads, this is one outcome, tails is one outcome. And this set, which has two elements, is the sample space of the experiment.

OK. What do we need to think about when we're setting up the sample space? First, the list should be mutually exclusive, collectively exhaustive. What does that mean?

Collectively exhaustive means that, no matter what happens in the experiment, you're going to get one of the outcomes inside here. So you have not forgotten any of the possibilities of what may happen in the experiment. Mutually exclusive means that if this happens, then that cannot happen. So at the end of the experiment, you should be able to point out to me just one, exactly one, of these outcomes and say, this is the outcome that happened.

OK. So these are sort of basic requirements. There's another requirement which is a little more loose. When you set up your sample space, sometimes you do have some freedom about the details of how you're going to describe it. And the question is, how much detail are you going to include?

So let's take this coin flipping experiment and think of the following sample space. One possible outcome is heads, a second possible outcome is tails and it's raining, and the third possible outcome is tails and it's not raining. So this is another possible sample space for the experiment where I flip a coin just once. It's a legitimate one. These three possibilities are mutually exclusive and collectively exhaustive.

Which one is the right sample space? Is it this one or that one? Well, if you think that my coin flipping inside this room is completely unrelated to the weather outside, then you're going to stick with this sample space. If, on the other hand, you have some superstitious belief that maybe

rain has an effect on my coins, you might work with the sample space of this kind. So you probably wouldn't do that, but it's a legitimate option, strictly speaking.

Now this example is a little bit on the frivolous side, but the issue that comes up here is a basic one that shows up anywhere in science and engineering. Whenever you're dealing with a model or with a situation, there are zillions of details in that situation. And when you come up with a model, you choose some of those details that you keep in your model, and some that you say, well, these are irrelevant. Or maybe there are small effects, I can neglect them, and you keep them outside your model. So when you go to the real world, there's definitely an element of art and some judgment that you need to do in order to set up an appropriate sample space.

So, an easy example now. So of course, the elementary examples are coins, cards, and dice. So let's deal with dice. But to keep the diagram small, instead of a six-sided die, we're going to think about the die that only has four faces. So you can do that with a tetrahedron, doesn't really matter. Basically, it's a die that when you roll it, you get a result which is one, two, three or four.

However, the experiment that I'm going to think about will consist of two rolls of a dice. A crucial point here-- I'm rolling the die twice, but I'm thinking of this as just one experiment, not two different experiments, not a repetition twice of the same experiment. So it's one big experiment. During that big experiment various things could happen, such as I'm rolling the die once, and then I'm rolling the die twice.

OK. So what's the sample space for that experiment? Well, the sample space consists of the possible outcomes. One possible outcome is that your first roll resulted in two and the second roll resulted in three. In which case, the outcome that you get is this one, a two followed by three. This is one possible outcome.

The way I'm describing things, this outcome is to be distinguished from this outcome here, where a three is followed by two. If you're playing backgammon, it doesn't matter which one of the two happened. But if you're dealing with a probabilistic model that you want to keep track of everything that happens in this composite experiment, there are good reasons for distinguishing between these two outcomes. I mean, when this happens, it's definitely something different from that happening. A two followed by a three is different from a three followed by a two.

So this is the correct sample space for this experiment where we roll the die twice. It has a total of 16 elements and it's, of course, a finite set.

Sometimes, instead of describing sample spaces in terms of lists, or sets, or diagrams of this kind, it's useful to describe the experiment in some sequential way. Whenever you have an experiment that consists of multiple stages, it might be useful, at least visually, to give a diagram that shows you how those stages evolve. And that's what we do by using a sequential description or a tree-based description by drawing a tree of the possible evolutions during our experiment.

So in this tree, I'm thinking of a first stage in which I roll the first die, and there are four possible results, one, two, three and four. And, given what happened, let's say in the first roll,

suppose I got a one. Then I'm rolling the second dice, and there are four possibilities for what may happen to the second die. And the possible results are one, two, three and four again.

So what's the relation between the two diagrams? Well, for example, the outcome two followed by three corresponds to this path on the tree. So this path corresponds to two followed by a three. Any path is associated to a particular outcome, any outcome is associated to a particular path.

And, instead of paths, you may want to think in terms of the leaves of this diagram. Same thing, think of each one of the leaves as being one possible outcome. And of course we have 16 outcomes here, we have 16 outcomes here.

Maybe you noticed the subtlety that I used in my language. I said I rolled the first dice and the result that I get is a two. I didn't use the word "outcome." I want to reserve the word "outcome" to mean the overall outcome at the end of the overall experiment.

So "2, 3" is the outcome of the experiment. The experiment consisted of stages. Two was the result in the first stage, three was the result in the second stage. You put all those results together, and you get your outcome. OK, perhaps we are splitting hairs here, but it's useful to keep the concepts right.

What's special about this example is that, besides being trivial, it has a sample space which is finite. There's 16 possible total outcomes. Not every experiment has a finite sample space.

Here's an experiment in which the sample space is infinite. So you are playing darts and the target is this square. And you're perfect at that game, so you're sure that your darts will always fall inside the square. So, but where exactly your dart would fall inside that square, that itself is random. We don't know what it's going to be. It's uncertain.

So all the possible points inside the square are possible outcomes of the experiment. So a typical outcome of the experiment is going to a pair of numbers, x, y , where x and y are real numbers between zero and one. Now there's infinitely many real numbers, there's infinitely many points in the square, so this is an example in which our sample space is an infinite set. OK, so we're going to revisit this example a little later.

So these are two examples of what the sample space might be in simple experiments. Now, the more important order of business is now to look at those possible outcomes and to make some statements about their relative likelihoods. Which outcome is more likely to occur compared to the others? And the way we do this is by assigning probabilities to the outcomes.

Well, not exactly. Suppose that all you were to do was to assign probabilities to individual outcomes. If you go back to this example, and you consider one particular outcome-- let's say this point-- what would be the probability that you hit exactly this point to infinite precision? Intuitively, that probability would be zero. So any individual point in this diagram in any reasonable model should have zero probability. So if you just tell me that any individual outcome has zero probability, you're not really telling me much to work with.

For that reason, what instead we're going to do is to assign probabilities to subsets of the sample space, as opposed to assigning probabilities to individual outcomes. So here's the picture. We have our sample space, which is omega, and we consider some subset of the sample space. Call it A. And I want to assign a number, a numerical probability, to this particular subset which represents my belief about how likely this set is to occur.

OK. What do we mean "to occur?" And I'm introducing here a language that's being used in probability theory. When we talk about subsets of the sample space, we usually call them events, as opposed to subsets. And the reason is because it works nicely with the language that describes what's going on.

So the outcome is a point. The outcome is random. The outcome may be inside this set, in which case we say that event A occurred, if we get an outcome inside here. Or the outcome may fall outside the set, in which case we say that event A did not occur.

So we're going to assign probabilities to events. And now, how should we do this assignment? Well, probabilities are meant to describe your beliefs about which sets are more likely to occur versus other sets. So there's many ways that you can assign those probabilities. But there are some ground rules for this game.

First, we want probabilities to be numbers between zero and one because that's the usual convention. So a probability of zero means we're certain that something is not going to happen. Probability of one means that we're essentially certain that something's going to happen. So we want numbers between zero and one.

We also want a few other things. And those few other things are going to be encapsulated in a set of axioms. What "axioms" means in this context, it's the ground rules that any legitimate probabilistic model should obey. You have a choice of what kind of probabilities you use. But, no matter what you use, they should obey certain consistency properties because if they obey those properties, then you can go ahead and do useful calculations and do some useful reasoning.

So what are these properties? First, probabilities should be non-negative. OK? That's our convention. We want probabilities to be numbers between zero and one. So they should certainly be non-negative. The probability that event A occurs should be a non-negative number.

What's the second axiom? The probability of the entire sample space is equal to one. Why does this make sense? Well, the outcome is certain to be an element of the sample space because we set up a sample space, which is collectively exhaustive. No matter what the outcome is, it's going to be an element of the sample space. We're certain that event omega is going to occur.

Therefore, we represent this certainty by saying that the probability of omega is equal to one.

Pretty straightforward so far. The more interesting axiom is the third rule. Before getting into it, just a quick reminder.

If you have two sets, A and B, the intersection of A and B consists of those elements that belong both to A and B. And we denote it this way. When you think probabilistically, the way to think

of intersection is by using the word "and." This event, this intersection, is the event that A occurred and B occurred. If I get an outcome inside here, A has occurred and B has occurred at the same time. So you may find the word "and" to be a little more convenient than the word "intersection."

And similarly, we have some notation for the union of two events, which we write this way. The union of two sets, or two events, is the collection of all the elements that belong either to the first set, or to the second, or to both. When you talk about events, you can use the word "or." So this is the event that A occurred or B occurred. And this "or" means that it could also be that both of them occurred.

OK. So now that we have this notation, what does the third axiom say? The third axiom says that if we have two events, A and B, that have no common elements-- so here's A, here's B, and perhaps this is our big sample space. The two events have no common elements. So the intersection of the two events is the empty set. There's nothing in their intersection. Then, the total probability of A together with B has to be equal to the sum of the individual probabilities. So the probability that A occurs or B occurs is equal to the probability that A occurs plus the probability that B occurs.

So think of probability as being cream cheese. You have one pound of cream cheese, the total probability assigned to the entire sample space. And that cream cheese is spread out over this set. The probability of A is how much cream cheese sits on top of A. Probability of B is how much sits on top of B. The probability of A union B is the total amount of cream cheese sitting on top of this and that, which is obviously the sum of how much is sitting here and how much is sitting there.

So probabilities behave like cream cheese, or they behave like mass. For example, if you think of some material object, the mass of this set consisting of two pieces is obviously the sum of the two masses. So this property is a very intuitive one. It's a pretty natural one to have.

OK. Are these axioms enough for what we want to do? I mentioned a while ago that we want probabilities to be numbers between zero and one. Here's an axiom that tells you that probabilities are non-negative. Should we have another axiom that tells us that probabilities are less than or equal to one? It's a desirable property. We would like to have it in our hands.

OK, why is it not in that list? Well, the people who are in the axiom making business are mathematicians and mathematicians tend to be pretty laconic. You don't say something if you don't have to say it. And this is the case here. We don't need that extra axiom because we can derive it from the existing axioms.

Here's how it goes. One is the probability over the entire sample space. Here we're using the second axiom. Now the sample space consists of A together with the complement of A. OK? When I write the complement of A, I mean the complement of A inside of the set omega. So we have omega, here's A, here's the complement of A, and the overall set is omega.

OK. Now, what's the next step? What should I do next? Which axiom should I use? We use axiom three because a set and the complement of that set are disjoint. They don't have any common elements. So axiom three applies and tells me that this is the probability of A plus the probability of A complement. In particular, the probability of A is equal to one minus the probability of A complement, and this is less than or equal to one.

Why? Because probabilities are non-negative, by the first axiom.

OK. So we got the conclusion that we wanted. Probabilities are always less than or equal to one, and this is a simple consequence of the three axioms that we have. This is a really nice argument because it actually uses each one of those axioms. The argument is simple, but you have to use all of these three properties to get the conclusion that you want.

OK. So we can get interesting things out of our axioms. Can we get some more interesting ones? How about the union of three sets? What kind of probability should it have?

So here's an event consisting of three pieces. And I want to say something about the probability of A union B union C. What I would like to say is that this probability is equal to the sum of the three individual probabilities. How can I do it?

I have an axiom that tells me that I can do it for two events. I don't have an axiom for three events. Well, maybe I can manage things and still be able to use that axiom. And here's the trick. The union of three sets, you can think of it as forming the union of the first two sets and then taking the union with the third set. OK? So taking unions, you can take the unions in any order that you want.

So here we have the union of two sets. Now, ABC are disjoint, by assumption or that's how I drew it. So if A, B, and C are disjoint, then A union B is disjoint from C. So here we have the union of two disjoint sets. So by the additivity axiom, the probability of that the union is going to be the probability of the first set plus the probability of the second set.

And now I can use the additivity axiom once more to write that this is probability of A plus probability of B plus probability of C. So by using this axiom which was stated for two sets, we can actually derive a similar property for the union of three disjoint sets. And then you can repeat this argument as many times as you want. It's valid for the union of ten disjoint sets, for the union of a hundred disjoint sets, for the union of any finite number of sets. So if A₁ up to A_n are disjoint, then the probability of A₁ union A_n is equal to the sum of the probabilities of the individual sets.

OK. Special case of this is when we're dealing with finite sets. Suppose I have just a finite set of outcomes. I put them together in a set and I'm interested in the probability of that set. So here's our sample space. There's lots of outcomes, but I'm taking a few of these and I form a set out of them.

This is a set consisting of, in this picture, three elements. In general, it consists of k elements. Now, a finite set, I can write it as a union of single element sets. So this set here is the union of

this one element set, together with this one element set together with that one element set. So the total probability of this set is going to be the sum of the probabilities of the one element sets.

Now, probability of a one element set, you need to use the brackets here because probabilities are assigned to sets. But this gets kind of tedious, so here one abuses notation a little bit and we get rid of those brackets and just write probability of this single, individual outcome. In any case, conclusion from this exercise is that the total probability of a finite collection of possible outcomes, the total probability is equal to the sum of the probabilities of individual elements.

So these are basically the axioms of probability theory. Or, well, they're almost the axioms. There are some subtleties that are involved here.

One subtlety is that this axiom here doesn't quite do the job for everything we would like to do. And we're going to come back to this at the end of the lecture. A second subtlety has to do with weird sets.

We said that an event is a subset of the sample space and we assign probabilities to events. Does this mean that we are going to assign probability to every possible subset of the sample space? Ideally, we would wish to do that. Unfortunately, this is not always possible.

If you take a sample space, such as the square, the square has nice subsets, those that you can describe by cutting it with lines and so on. But it does have some very ugly subsets, as well, that are impossible to visualize, impossible to imagine, but they do exist. And those very weird sets are such that there's no way to assign probabilities to them in a way that's consistent with the axioms of probability.

OK. So this is a very, very fine point that you can immediately forget for the rest of this class. You will only encounter these sets if you end up doing doctoral work on the theoretical aspects of probability theory. So it's just a mathematical subtlety that some very weird sets do not have probabilities assigned to them. But we're not going to encounter these sets and they do not show up in any applications.

OK. So now let's revisit our examples. Let's go back to the die example. We have our sample space. Now we need to assign a probability law. There's lots of possible probability laws that you can assign. I'm picking one here, arbitrarily, in which I say that every possible outcome has the same probability of 1/16.

OK. Why do I make this model? Well, empirically, if you have well-manufactured dice, they tend to behave that way. We will be coming back to this kind of story later in this class. But I'm not saying that this is the only probability law that there can be. You might have weird dice in which certain outcomes are more likely than others. But to keep things simple, let's take every outcome to have the same probability of 1/16.

OK. Now that we have in our hands a sample space and the probability law, we can actually solve any problem there is. We can answer any question that could be posed to us. For example,

what's the probability that the outcome, which is this pair, is either 1,1 or 1,2. We're talking here about this particular event, 1,1 or 1,2. So it's an event consisting of these two items.

According to what we were just discussing, the probability of a finite collection of outcomes is the sum of their individual probabilities. Each one of them has probability of 1/16, so the probability of this is 2/16.

How about the probability of the event that x is equal to one. x is the first roll, so that's the probability that the first roll is equal to one. Notice the syntax that's being used here. Probabilities are assigned to subsets, to sets, so we think of this as meaning the set of all outcomes such that x is equal to one.

How do you answer this question? You go back to the picture and you try to visualize or identify this event of interest. x is equal to one corresponds to this event here. These are all the outcomes at which x is equal to one. There's four outcomes. Each one has probability 1/16, so the answer is 4/16.

OK. How about the probability that x plus y is odd? OK. That will take a little bit more work.

But you go to the sample space and you identify all the outcomes at which the sum is an odd number. So that's a place where the sum is odd, these are other places, and I guess that exhausts all the possible outcomes at which we have an odd sum. We count them. How many are there? There's a total of eight of them. Each one has probability 1/16, total probability is 8/16.

And harder question. What is the probability that the minimum of the two rolls is equal to 2? This is something that you probably couldn't do in your head without the help of a diagram. But once you have a diagram, things are simple.

You ask the question. OK, this is an event, that the minimum of the two rolls is equal to two. This can happen in several ways. What are the several ways that it can happen? Go to the diagram and try to identify them.

So the minimum is equal to two if both of them are two's. Or it could be that x is two and y is bigger, or y is two and x is bigger. OK. I guess we rediscover that yellow and blue make green, so we see here that there's a total of five possible outcomes. The probability of this event is 5/16.

Simple example, but the procedure that we followed in this example actually applies to any probability model you might ever encounter. You set up your sample space, you make a statement that describes the probability law over that sample space, then somebody asks you questions about various events. You go to your pictures, identify those events, pin them down, and then start kind of counting and calculating the total probability for those outcomes that you're considering.

This example is a special case of what is called the discrete uniform law. The model obeys the discrete uniform law if all outcomes are equally likely. It doesn't have to be that way. That's just one example of a probability law.

But when things are that way, if all outcomes are equally likely and we have N of them, and you have a set A that has little n elements, then each one of those elements has probability one over capital N since all outcomes are equally likely. And for our probabilities to add up to one, each one must have this much probability, and there's little n elements. That gives you the probability of the event of interest.

So problems like the one in the previous slide and more generally of the type described here under discrete uniform law, these problems reduce to just counting. How many elements are there in my sample space? How many elements are there inside the event of interest? Counting is generally simple, but for some problems it gets pretty complicated. And in a couple of weeks, we're going to have to spend the whole lecture just on the subject of how to count systematically.

Now the procedure we followed in the previous example is the same as the procedure you would follow in continuous probability problems. So, going back to our dart problem, we get the random point inside the square. That's our sample space. We need to assign a probability law. For lack of imagination, I'm taking the probability law to be the area of a subset.

So if we have two subsets of the sample space that have equal areas, then I'm postulating that they are equally likely to occur. The probably that they fall here is the same as the probability that they fall there. The model doesn't have to be that way. But if I have sort of complete ignorance of which points are more likely than others, that might be the reasonable model to use.

So equal areas mean equal probabilities. If the area is twice as large, the probability is going to be twice as big. So this is our model.

We can now answer questions. Let's answer the easy one. What's the probability that the outcome is exactly this point? That of course is zero because a single point has zero area. And since this probability is equal to area, that's zero probability.

How about the probability that the sum of the coordinates of the point that we got is less than or equal to $1/2$? How do you deal with it? Well, you look at the picture again, at your sample space, and try to describe the event that you're talking about. The sum being less than $1/2$ corresponds to getting an outcome that's below this line, where this line is the line where x plus y equals to $1/2$. So the intercepts of that line with the axis are $1/2$ and $1/2$.

So you describe the event visually and then you use your probability law. The probability law that we have is that the probability of a set is equal to the area of that set. So all we need to find is the area of this triangle, which is $1/2$ times $1/2$ times $1/2$, half, equals to $1/8$.

OK. Moral from these two examples is that it's always useful to have a picture and work with a picture to visualize the events that you're talking about. And once you have a probability law in your hands, then it's a matter of calculation to find the probabilities of an event of interest. The calculations we did in these two examples, of course, were very simple.

Sometimes calculations may be a lot harder, but it's a different business. It's a business of calculus, for example, or being good in algebra and so on. As far as probability is concerned, it's

clear what you will be doing, and then maybe you're faced with a harder algebraic part to actually carry out the calculations. The area of a triangle is easy to compute. If I had put down a very complicated shape, then you might need to solve a hard integration problem to find the area of that shape, but that's stuff that belongs to another class that you have presumably mastered by now.

Good, OK. So now let me spend just a couple of minutes to return to a point that I raised before. I was saying that the axiom that we had about additivity might not quite be enough. Let's illustrate what I mean by the following example.

Think of the experiment where you keep flipping a coin and you wait until you obtain heads for the first time. What's the sample space of this experiment? It might happen the first flip, it might happen in the tenth flip. Heads for the first time might occur in the millionth flip.

So the outcome of this experiment is going to be an integer and there's no bound to that integer. You might have to wait very much until that happens. So the natural sample space is the set of all possible integers.

Somebody tells you some information about the probability law. The probability that you have to wait for n flips is equal to two to the minus n . Where did this come from? That's a separate story. Where did it come from? Somebody tells this to us, and those probabilities are plotted here as a function of n .

And you're asked to find the probability that the outcome is an even number. How do you go about calculating that probability? So the probability of being an even number is the probability of the subset that consists of just the even numbers. So it would be a subset of this kind, that includes two, four, and so on.

So any reasonable person would say, well the probability of obtaining an outcome that's either two or four or six and so on is equal to the probability of obtaining a two, plus the probability of obtaining a four, plus the probability of obtaining a six, and so on. These probabilities are given to us. So here I have to do my algebra. I add this geometric series and I get an answer of $1/3$. That's what any reasonable person would do.

But the person who only knows the axioms that they posted just a little earlier may get stuck. They would get stuck at this point. How do we justify this?

We had this property for the union of disjoint sets and the corresponding property that tells us that the total probability of finitely many things, outcomes, is the sum of their individual probabilities. But here we're using it on an infinite collection. The probability of infinitely many points is equal to the sum of the probabilities of each one of these. To justify this step we need to introduce one additional rule, an additional axiom, that tells us that this step is actually legitimate.

And this is the countable additivity axiom, which is a little stronger, or quite a bit stronger, than the additivity axiom we had before. It tells us that if we have a sequence of sets that are disjoint

and we want to find their total probability, then we are allowed to add their individual probabilities. So the picture might be such as follows.

We have a sequence of sets, A_1 , A_2 , A_3 , and so on. I guess in order to fit them inside the sample space, the sets need to get smaller and smaller perhaps. They are disjoint. We have a sequence of such sets. The total probability of falling anywhere inside one of those sets is the sum of their individual probabilities.

A key subtlety that's involved here is that we're talking about a sequence of events. By "sequence" we mean that these events can be arranged in order. I can tell you the first event, the second event, the third event, and so on. So if you have such a collection of events that can be ordered as first, second, third, and so on, then you can add their probabilities to find the probability of their union.

So this point is actually a little more subtle than you might appreciate at this point, and I'm going to return to it at the beginning of the next lecture. For now, enjoy the first week of classes and have a good weekend. Thank you.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013

Transcript – Recitation: Uniform Probabilities on a Square

In this problem, we will be helping Romeo and Juliet meet up for a date. And in the process, also we'll review some concepts in basic probability theory, including sample spaces and probability laws.

This problem, the basic setup is that Romeo and Juliet are trying to meet up for a date. And let's say they're trying to meet up for lunch tomorrow at noon. But they're not necessarily punctual. So they may arrive on time with a delay of 0, or they may actually be up to 1 hour late and arrive at 1:00 PM.

So the other thing that we assume in this problem is that all pairs of arrival times-- so the time that Romeo arrives paired with the time they Juliet arrives-- all of these pairs are equally likely. And I've put this in quotes, because we haven't really specify exactly what this means. And we'll come back to that in a little bit. The last important thing is that each person will wait for 15 minutes for the other person to arrive. If within that 15-minute window the other person doesn't arrive, then they'll give up and they'll end up not meeting up for lunch.

So to solve this problem, let's first try to set up a sample space and come up with a probability law to describe this scenario. And let's actually start with a simpler version of this problem. And instead of assuming that they can arrive at any delay between 0 and 1 hour, let's pretend instead that Romeo and Juliet can only arrive in 15-minute increments. So Romeo can arrive on time with a delay 0, or be 15 minutes late, 30 minutes late, 45 minutes late, or one hour late. But none of the other times are possible. And the same thing for Juliet.

Let's start out with just the simple case first, because it helps us get the intuition for the problem, and it's an easier case to analyze. So it's actually easy to visualize this. It's a nice visual tool to group this sample space into a grid. So the horizontal axis here represents the arrival time of Romeo, and the vertical axis represents the arrival time of Juliet. And so, for example, this point here would represent Romeo arriving 15 minutes late and Juliet arriving 30 minutes late.

So this is our sample space now. This is our Ω . And now let's try to assign a probability law. And we'll continue to assume that all pairs of arrival times are equally likely. And now we can actually specifically specify what this term means.

And in particular, we'll be invoking the discrete uniform law, which basically says that all of these points, which are just outcomes in our probabilistic experiment-- all of these outcomes are equally likely. And so since there are 25 of them, each one of these outcomes has a probability of 1 over 25.

So now we've specified our sample space and our probability law. So now let's try to answer the question, what is the probability that Romeo and Juliet will meet up for their date? So all that amounts to now is just identifying which of these 25 outcomes results in Romeo and Juliet arriving within 15 minutes of each other.

So let's start with this one that I've picked out. If Romeo arrives 15 minutes late and Juliet arrives 30 minutes late, then they will arrive within 15 minutes of each other. So this outcome does result in the two of them meeting. And so we can actually highlight all of these. And it turns out that these outcomes that I'm highlighting result in the two them arriving within 15 minutes of each other.

So because each one has a probability of 1 over 25, all we really need to do now is just count how many outcomes there are. So there's 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13. So the probability in the end is for the discrete case.

The discrete case-- I'm referring to the case where we simplified it and considered only arrival times with increments of 15 minutes. In this case, the probability is 13 over 25. So now we have an idea of how to solve this problem. It amounts to basically coming up with a sample space, a probability law, and then identifying the events of interest in calculating the probability of that event.

So now let's actually solve the problem that we really are interested in, which is that instead of confining Romeo and Juliet to arrive in only 15-minute minute increments, really, time is continuous, and Romeo and Juliet can arrive at any time. So they don't necessarily have to arrive 15 minutes late. Romeo could arrive 15 minutes and 37 seconds late if he wanted to.

So now our new sample space is actually just, instead of only these 25 points in the grid, it's this entire square. So any point within the square could be a possible pair of meeting times between Romeo and Juliet. So that is our new sample space, our new omega.

And now let's assign a new probability law. And now, instead of being in the discrete world, we're in the continuous world. And the analogy here is to consider probabilities as areas. So the area of this entire square is one. And that also corresponds to the probability of omega, the sample space. And imagine just spreading probability evenly across this square so that the probability of any event-- which in this case would just be any shape within this square-- is exactly equal to the area of that shape. So now that is our new sample space and our new probability law.

So what we have to do now is just to identify the event of interest, which is still the event that Romeo and Juliet arrive within 15 minutes of each other. So let's do that. If Romeo and Juliet arrive both on time, then obviously they'll meet. And if Romeo's on time and Juliet is 15 minutes late, then they will still meet.

And in fact, any pairs of meeting times between these would still work, because now Romeo can be on time, and Juliet can arrive at any time between 0 and 15 minutes late. But you notice that if Juliet is even a tiny bit later than 15 minutes, then they won't end up meeting. So this segment here is part of the event of interest.

And similarly, this segment here is also part of the event. And if you take this exercise and extend it, you can actually verify that the event of interest is this strip shape in the middle of the square. Which, if you think about it, makes sense, because you want the arrival times between

Romeo and Juliet to be close to each other, so you would expect it to be somewhere close to a diagonal in this square.

So now we have our event of interest. We have our sample space and our probability law. So all we have to do now is just calculate what this probability is. And we've already said that the probability in this probability law is just areas. So now it actually just boils down to not a probability problem, but a problem in geometry.

So to calculate this area, you can do it in lots of ways. One way is to calculate the area of the square, which is 1, and subtract the areas of these two triangles. So let's do that.

So in the continuous case, the probability of meeting is going to be 1 minus the area of this triangle. The base here is $3/4$ and $3/4$, so it's $1/2$ times $3/4$ times $3/4$. That's the area of one of these triangles. There's two of them, so we'll multiply by two. And we end up with $1 - 9/16$, or $7/16$ as our final answer.

So in this problem, we've reviewed some basic concepts of probability, and that's also helped us solve this problem of helping Romeo and Juliet meet up for a date. And if you wanted to, you could even extend this problem even further and turn it on its head. And instead of calculating given that they arrive within 15 minutes of each other, what is the probability that they'll meet, let's say that Romeo really wants to meet up with Juliet, and he wants to assure himself at least, say, a 90% chance of meeting Juliet. Then you can ask, if he wants to have at least a 90% chance of meeting her, how long should he be willing to wait? And so that's the flip side of the problem.

And you can see that with just some basic concepts of probability, you can answer some already pretty interesting problems. So I hope this problem was interesting, and we'll see you next time.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Spring 2010)

Problem Set 1
Due: September 15, 2010

1. Express each of the following events in terms of the events A , B and C as well as the operations of complementation, union and intersection:
 - (a) at least one of the events A , B , C occurs;
 - (b) at most one of the events A , B , C occurs;
 - (c) none of the events A , B , C occurs;
 - (d) all three events A , B , C occur;
 - (e) exactly one of the events A , B , C occurs;
 - (f) events A and B occur, but not C ;
 - (g) either event A occurs or, if not, then B also does not occur.

In each case draw the corresponding Venn diagrams.

2. You flip a fair coin 3 times, determine the probability of the below events. Assume all sequences are equally likely.
 - (a) Three heads: HHH
 - (b) The sequence head, tail, head: HTH
 - (c) Any sequence with 2 heads and 1 tail
 - (d) Any sequence where the number of heads is greater than or equal to the number of tails
3. Bob has a peculiar pair of four-sided dice. When he rolls the dice, the probability of any particular outcome is proportional to the sum of the results of each die. All outcomes that result in a particular sum are equally likely.
 - (a) What is the probability of the sum being even?
 - (b) What is the probability of Bob rolling a 2 and a 3, in any order?
4. Alice and Bob each choose at random a number in the interval $[0, 2]$. We assume a uniform probability law under which the probability of an event is proportional to its area. Consider the following events:

A : The magnitude of the difference of the two numbers is greater than $1/3$.

B : At least one of the numbers is greater than $1/3$.

C : The two numbers are equal.

D : Alice's number is greater than $1/3$.

Find the probabilities $\mathbf{P}(B)$, $\mathbf{P}(C)$, and $\mathbf{P}(A \cap D)$.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Spring 2010)

5. Mike and John are playing a friendly game of darts where the dart board is a disk with radius of 10in.

Whenever a dart falls within 1in of the center, 50 points are scored. If the point of impact is between 1 and 3in from the center, 30 points are scored, if it is at a distance of 3 to 5in 20 points are scored and if it is further than 5in, 10 points are scored.

Assume that both players are skilled enough to be able to throw the dart within the boundaries of the board.

Mike can place the dart uniformly on the board (i.e., the probability of the dart falling in a given region is proportional to its area).

- (a) What is the probability that Mike scores 50 points on one throw?
 - (b) What is the probability of him scoring 30 points on one throw?
 - (c) John is right handed and is twice more likely to throw in the right half of the board than in the left half. Across each half, the dart falls uniformly in that region. Answer the previous questions for John's throw.
6. Prove that for any three events A , B and C , we have

$$\mathbf{P}(A \cap B \cap C) \geq \mathbf{P}(A) + \mathbf{P}(B) + \mathbf{P}(C) - 2.$$

G1[†]. Consider an experiment whose sample space is the real line.

- (a) Let $\{a_n\}$ be an increasing sequence of numbers that converges to a and $\{b_n\}$ a decreasing sequence that converges to b . Show that

$$\lim_{n \rightarrow \infty} \mathbf{P}([a_n, b_n]) = \mathbf{P}([a, b]).$$

Here, the notation $[a, b]$ stands for the closed interval $\{x \mid a \leq x \leq b\}$. *Note:* This result seems intuitively obvious. The issue is to derive it using the axioms of probability theory.

- (b) Let $\{a_n\}$ be a decreasing sequence that converges to a and $\{b_n\}$ an increasing sequence that converges to b . Is it true that

$$\lim_{n \rightarrow \infty} \mathbf{P}([a_n, b_n]) = \mathbf{P}([a, b])?$$

Note: You may use freely the results from the problems in the text in your proofs.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

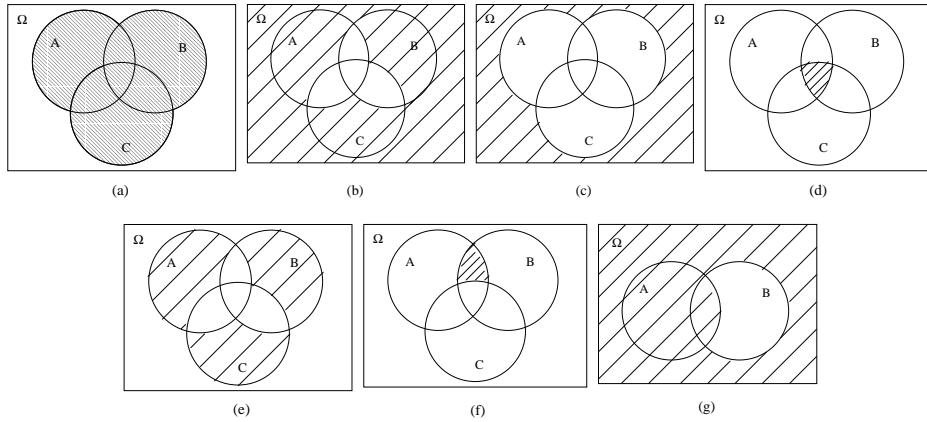
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
 (Fall 2010)

Problem Set 1: Solutions
Due: September 15, 2010

1. (a) $A \cup B \cup C$
- (b) $(A \cap B^c \cap C^c) \cup (A^c \cap B \cap C^c) \cup (A^c \cap B^c \cap C) \cup (A^c \cap B^c \cap C^c)$
- (c) $(A \cup B \cup C)^c = A^c \cap B^c \cap C^c$
- (d) $A \cap B \cap C$
- (e) $(A \cap B^c \cap C^c) \cup (A^c \cap B \cap C^c) \cup (A^c \cap B^c \cap C)$
- (f) $A \cap B \cap C^c$
- (g) $A \cup (A^c \cap B^c)$



2. Since all outcomes are equally likely we apply the discrete uniform probability law to solve the problem. To solve for any event we simply count the number of elements in the event and divide by the total number of elements in the sample space.

There are 2 possible outcomes for each flip, and 3 flips. Thus there are $2^3 = 8$ elements (or sequences) in the sample space.

- (a) Any sequence has probability of $1/8$. Therefore $\mathbf{P}(\{H, H, H\}) = \boxed{1/8}$.
- (b) This is still a single sequence, thus $\mathbf{P}(\{H, T, H\}) = \boxed{1/8}$.
- (c) The event of interest has 3 unique sequences, thus $\mathbf{P}(\{HHT, HTH, THH\}) = \boxed{3/8}$.
- (d) The sequences where there are more heads than tails are $A : \{HHH, HHT, HTH, THH\}$.
 4 unique sequences gives us $\mathbf{P}(A) = \boxed{1/2}$.
3. The easiest way to solve this problem is to make a table of some sort, similar to the one below.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
 (Fall 2010)

Die 1	Die 2	Sum	P(Sum)
1	1	2	2p
1	2	3	3p
1	3	4	4p
1	4	5	5p
2	1	3	3p
2	2	4	4p
2	3	5	5p
2	4	6	6p
3	1	4	4p
3	2	5	5p
3	3	6	6p
3	4	7	7p
4	1	5	5p
4	2	6	6p
4	3	7	7p
4	4	8	8p
Total		80p	

$$\mathbf{P}(\text{All outcomes}) = 80p \text{ (Total from the table)}$$

and therefore

$$p = \frac{1}{80}$$

(a)

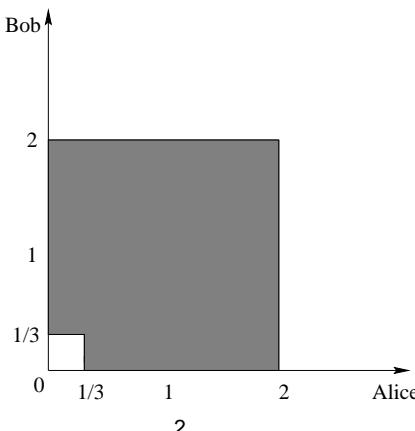
$$\mathbf{P}(\text{Even sum}) = 2p + 4p + 4p + 6p + 4p + 6p + 6p + 8p = 40p = \boxed{1/2}$$

(b)

$$\mathbf{P}(\text{Rolling a 2 and a 3}) = \mathbf{P}(2,3) + \mathbf{P}(3,2) = 5p + 5p = 10p = \boxed{1/8}$$

4. $\mathbf{P(B)}$

The shaded area in the following figure is the union of Alice's pick being greater than $1/3$ and Bob's pick being greater than $1/3$.

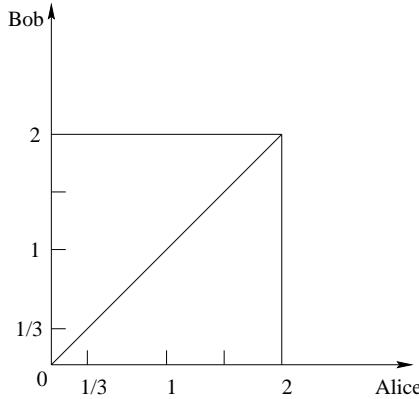


MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
 (Fall 2010)

$$\begin{aligned}
 \mathbf{P}(B) &= 1 - \mathbf{P}(\text{both numbers are smaller than } 1/3) \\
 &= 1 - \frac{\text{area of small square}}{\text{total sample area}} \\
 &= 1 - \frac{(1/3)(1/3)}{4} = 1 - \frac{1}{36} = \boxed{35/36}
 \end{aligned}$$

P(C)

In the following figure, the diagonal line represents the set of points where the two selected numbers are equal.

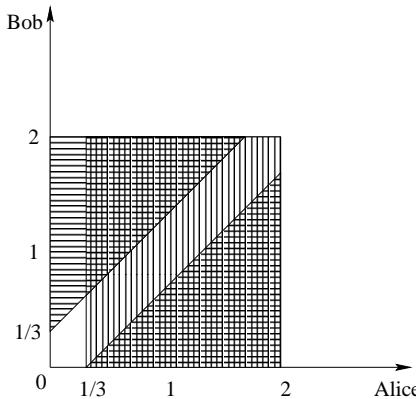


The line has an area of 0. Thus,

$$\mathbf{P}(C) = \frac{\text{area of line}}{\text{total sample area}} = \frac{0}{4} = \boxed{0}$$

P(A ∩ D)

Overlapping the diagrams we would get for **P(A)** and **P(D)**,



MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
 (Fall 2010)

$$\begin{aligned}\mathbf{P}(A \cap D) &= \frac{\text{double shaded area}}{\text{total sample area}} \\ &= \frac{(5/3)(5/3)(1/2) + (4/3)(4/3)(1/2)}{4} = \frac{25/18 + 16/18}{4} = \boxed{41/72}\end{aligned}$$

5. (a) The probability of Mike scoring 50 points is proportional to the area of the inner disk. Hence, it is equal to $\alpha\pi R^2 = \alpha\pi$, where α is a constant to be determined.

Since the probability of landing the dart on the board is equal to one, $\alpha\pi 10^2 = 1$, which implies that $\alpha = 1/(100\pi)$.

Therefore, the probability that Mike scores 50 points is equal to $\pi/(100\pi) = \boxed{0.01}$

- (b) In order to score exactly 30 points, Mike needs to place the dart between 1 and 3 inches from the origin. An easy way to compute this probability is to look first at that of scoring *more* than 30 points, which is equal to $\alpha\pi 3^2 = 0.09$.

Next, since the 30 points ring is disjoint from the 50 points disc, probability of scoring more than 30 points is equal to the probability of scoring 50 points plus that of scoring exactly 30 points. Hence, the probability of Mike scoring exactly 30 points is equal to $0.09 - 0.01 = \boxed{0.08}$

- (c) For the part (a) question. The probability of John scoring 50 points is equal to the probability of throwing in the right half of the board and scoring 50 points plus that of throwing in the left half and scoring 50 points.

The first term in the sum is proportional to the area of the right half of the inner disk and is equal to $\alpha\pi R^2/2 = \alpha\pi/2$, where α is a constant to be determined.

Similarly, the probability of him throwing in the left half of the board and scoring 50 points is equal to $\beta\pi/2$, where β is a constant (not necessarily equal to α).

In order to determine α and β , let us compute the probability of throwing the dart in the right half of the board. This probability is equal to

$$\alpha\pi R^2/2 = \alpha\pi 10^2/2 = \alpha 50\pi.$$

Since that probability is equal to $2/3$, $\alpha = 1/(75\pi)$. In a similar fashion, β can be determined to be $1/(150\pi)$. Consequently, the total probability is equal to $1/150 + 1/300 = \boxed{0.01}$

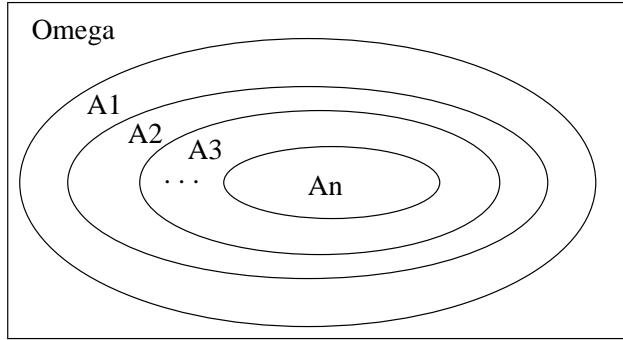
For the part (b), The probability of scoring exactly 30 points is equal to that of scoring more than 30 points minus that of scoring exactly 50. By applying the same type of analysis as in (b) above, the probability is found to be equal to $\boxed{0.08}$

These numbers suggest that John and Mike have similar skills, and are equally likely to win the game. The fact that Mike's better control (or worst, depending on how you look at it) of the direction of his throw does not increase his chances of winning can be explained by the observation that both players' control over the distance from the origin is identical.

6. See the textbook, Problem 1.11 page 55, which proves the general version of Bonferroni's inequality.

- G1[†]. (a) If we define $A_n = [a_n, b_n]$ for all n , it is easy to see that the sequence A_1, A_2, \dots is "monotonically decreasing," i.e., $A_{n+1} \subset A_n$ for all n :

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
 (Fall 2010)



Furthermore, $\cap_n^\infty A_n = [a, b]$.

By the continuity property of probabilities (see Problem 1.13, page 56 of the text),

$$\lim_{n \rightarrow \infty} \mathbf{P}([a_n, b_n]) = \mathbf{P}([a, b]).$$

- (b) No. Consider the following example. Let $a_n = a + \frac{1}{n}$, $b_n = b - \frac{1}{n}$ for all n . Then $\{a_n\}$ is a decreasing sequence that converges to a , and $\{b_n\}$ is an increasing sequence that converges to b . If we define a probability law that places non-zero probability only on points a and b , then $\lim_{n \rightarrow \infty} \mathbf{P}([a_n, b_n]) = 0$, but $\mathbf{P}([a, b]) = 1$.

This example is closely related to the continuity property of probabilities. In this case, if we define $A_n = [a_n, b_n]$, then A_1, A_2, \dots is “monotonically increasing,” i.e., $A_n \subset A_{n+1}$, but $A = (\cup_n^\infty A_n) = (a, b)$, which is an open interval whose probability is 0 under our probability law.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

LECTURE 1

- **Readings:** Sections 1.1, 1.2

Lecture outline

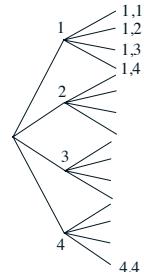
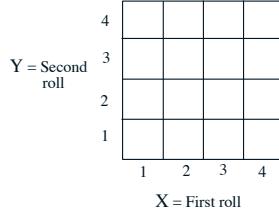
- Probability as a mathematical framework for:
 - reasoning about uncertainty
 - developing approaches to inference problems
- Probabilistic models
 - sample space
 - probability law
- Axioms of probability
- Simple examples

Sample space Ω

- “List” (set) of possible outcomes
- List must be:
 - Mutually exclusive
 - Collectively exhaustive
- Art: to be at the “right” granularity

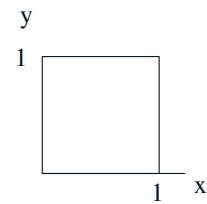
Sample space: Discrete example

- Two rolls of a tetrahedral die
 - Sample space vs. sequential description



Sample space: Continuous example

$$\Omega = \{(x, y) \mid 0 \leq x, y \leq 1\}$$



Probability axioms

- **Event:** a subset of the sample space
- Probability is assigned to events

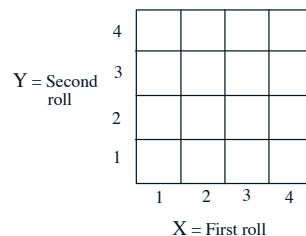
Axioms:

1. **Nonnegativity:** $P(A) \geq 0$
2. **Normalization:** $P(\Omega) = 1$
3. **Additivity:** If $A \cap B = \emptyset$, then $P(A \cup B) = P(A) + P(B)$

- $P(\{s_1, s_2, \dots, s_k\}) = P(\{s_1\}) + \dots + P(\{s_k\})$
 $= P(s_1) + \dots + P(s_k)$

- Axiom 3 needs strengthening
- Do weird sets have probabilities?

Probability law: Example with finite sample space



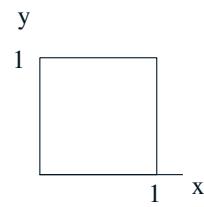
- Let every possible outcome have probability $1/16$
 - $P((X, Y) \text{ is } (1,1) \text{ or } (1,2)) =$
 - $P(\{X = 1\}) =$
 - $P(X + Y \text{ is odd}) =$
 - $P(\min(X, Y) = 2) =$

Discrete uniform law

- Let all outcomes be equally likely
- Then,
$$P(A) = \frac{\text{number of elements of } A}{\text{total number of sample points}}$$
- Computing probabilities \equiv counting
- Defines fair coins, fair dice, well-shuffled card decks

Continuous uniform law

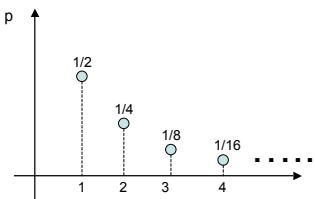
- Two “random” numbers in $[0, 1]$.



- **Uniform law:** Probability = Area
 - $P(X + Y \leq 1/2) = ?$
 - $P((X, Y) = (0.5, 0.3))$

Probability law: Ex. w/countably infinite sample space

- Sample space: $\{1, 2, \dots\}$
- We are given $P(n) = 2^{-n}$, $n = 1, 2, \dots$
- Find $P(\text{outcome is even})$



$$P(\{2, 4, 6, \dots\}) = P(2) + P(4) + \dots = \frac{1}{2^2} + \frac{1}{2^4} + \frac{1}{2^6} + \dots$$

- Countable additivity axiom (needed for this calculation):

If A_1, A_2, \dots are disjoint events, then:

$$P(A_1 \cup A_2 \cup \dots) = P(A_1) + P(A_2) + \dots$$

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Fall 2010)

Recitation 1
September 9, 2010

1. Give a mathematical derivation of the formula

$$\mathbf{P}((A \cap B^c) \cup (A^c \cap B)) = \mathbf{P}(A) + \mathbf{P}(B) - 2\mathbf{P}(A \cap B).$$

Your derivation should be a sequence of steps, with each step justified by appealing to one of the probability axioms.

2. Problem 1.5, page 54 in the text.

Out of the students in a class, 60% are geniuses, 70% love chocolate, and 40% fall into both categories. Determine the probability that a randomly selected student is neither a genius nor a chocolate lover.

3. A six-sided die is loaded in a way that each even face is twice as likely as each odd face. Construct a probabilistic model for a single roll of this die, and find the probability that a 1, 2, or 3 will come up.
4. Example 1.5, page 13 in the text.

Romeo and Juliet have a date at a given time, and each will arrive at the meeting place with a delay between 0 and 1 hour, with all pairs of delays being equally likely. The first to arrive will wait for 15 minutes and will leave if the other has not yet arrived. What is the probability that they will meet?

- G1[†]. Problem 1.13, page 56 in the text. **Continuity property of probabilities.**

- Let A_1, A_2, \dots be an infinite sequence of events that is “monotonically increasing,” meaning that $A_n \subset A_{n+1}$ for every n . Let $A = \bigcup_{n=1}^{\infty} A_n$. Show that $\mathbf{P}(A) = \lim_{n \rightarrow \infty} \mathbf{P}(A_n)$. *Hint:* Express the event A as a union of countably many disjoint sets.
- Suppose now that the events are “monotonically decreasing,” i.e., $A_{n+1} \subset A_n$ for every n . Let $A = \bigcap_{n=1}^{\infty} A_n$. Show that $\mathbf{P}(A) = \lim_{n \rightarrow \infty} \mathbf{P}(A_n)$. *Hint:* Apply the result of the previous part to the complements of the events.
- Consider a probabilistic model whose sample space is the real line. Show that

$$\mathbf{P}([0, \infty)) = \lim_{n \rightarrow \infty} \mathbf{P}([0, n]) \quad \text{and} \quad \lim_{n \rightarrow \infty} \mathbf{P}([n, \infty)) = 0.$$

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
 (Fall 2010)

Recitation 1: Solutions
September 9, 2010

1. Since the events $A \cap B^c$ and $A^c \cap B$ are disjoint, we have, using the additivity axiom,

$$\mathbf{P}((A \cap B^c) \cup (A^c \cap B)) = \mathbf{P}(A \cap B^c) + \mathbf{P}(A^c \cap B).$$

Since $A = (A \cap B) \cup (A \cap B^c)$ is the union of two disjoint sets, we have, again by the additivity axiom,

$$\mathbf{P}(A) = \mathbf{P}(A \cap B) + \mathbf{P}(A \cap B^c),$$

so that

$$\mathbf{P}(A \cap B^c) = \mathbf{P}(A) - \mathbf{P}(A \cap B).$$

Similarly,

$$\mathbf{P}(B \cap A^c) = \mathbf{P}(B) - \mathbf{P}(A \cap B).$$

Therefore,

$$\begin{aligned}\mathbf{P}(A \cap B^c) + \mathbf{P}(A^c \cap B) &= \mathbf{P}(A) - \mathbf{P}(A \cap B) + \mathbf{P}(B) - \mathbf{P}(A \cap B) \\ &= \mathbf{P}(A) + \mathbf{P}(B) - 2\mathbf{P}(A \cap B).\end{aligned}$$

2. Let

A : The event that the randomly selected student is a genius.

B : The event that the randomly selected student loves chocolate.

From the properties of probability laws proved in lecture, we have

$$\begin{aligned}1 &= \mathbf{P}(A \cup B) + \mathbf{P}((A \cup B)^c) \\ &= \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B) + \mathbf{P}(A^c \cap B^c) \\ &= 0.6 + 0.7 - 0.4 + \mathbf{P}(A^c \cap B^c) \\ &= 0.9 + \mathbf{P}(A^c \cap B^c).\end{aligned}$$

Therefore

$$\begin{aligned}\mathbf{P}(\text{A randomly selected student is neither a genius nor a chocolate lover}) \\ = \mathbf{P}(A^c \cap B^c) = 1 - 0.9 = 0.1.\end{aligned}$$

3. Let c denote the probability of a single odd face. Then the probability of a single even face is $2c$, and by adding the probabilities of the 3 odd faces and the 3 even faces, we get $9c = 1$. Thus, $c = 1/9$. The desired probability is

$$\mathbf{P}(\{1, 2, 3\}) = \mathbf{P}(\{1\}) + \mathbf{P}(\{2\}) + \mathbf{P}(\{3\}) = c + 2c + c = 4c = 4/9.$$

4. See the textbook, Example 1.5, page 13.

G1[†]. See the textbook, Problem 1.13, page 56.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013 Transcript – Recitation: A Coin Tossing Puzzle

Hi. In this problem, we'll be going over practice with the calculation of conditional probabilities. We'll start with a game where our friend Alice will be tossing a coin with certain bias of having a head, and tosses this coin twice. And we're interested in knowing, what's the probability that both coin tosses will end up being a head?

The first step we're going to do is to convert the problem into a mathematical form by defining two events as the following. Event A is where the first coin toss is a head. And similarly, event B will be having the second coin toss also being a head. Having these two events will allow us to say, well, the event that A intersection B will be the event that both coin tosses are a head. And we'd like to know the probability of such an event.

In particular, the probability of A and B will be calculated under two types of information. In the first case, we'll be conditioning on that we know the first coin toss is a head. I'd like to know what the probability of A and B is. In the second case, we know that at least one of the two coin tosses is a head expressed in the form A union B.

And under this conditioning, what is the probability of A and B, A intersection B? So Alice, in this problem, says-- well, her guess will be that the first quantity is no smaller than the second quantity. Namely, knowing that the first coin toss is a head somehow more strongly implies that both coin tosses will be a head, compared to the case that we only know at least one of the two coin tosses is a head.

And we'd like to verify if this inequality is indeed true. To do so, let's just use the basic calculation of conditional probability. Now, from the lectures, you've already learned that to calculate this quantity, we'll write out a fraction where the numerator is the probability of the intersection of these two events. So we have A intersect B intersection A divided by the probability of the event that we're conditioning on, which is A.

Now, the top quantity, since we know that A and B is a subset of event A, then taking the intersection of these two quantities will just give us the first event. So we have A and B. And the bottom is still probability of A.

Let's do the same thing for the second quantity here. We have the top probability of A and B intersection the event A union B, and on the bottom, probability of the event A and B. Again, we see the event A and B is a subset of the event A union B. So the top will be A and B. And the bottom-- A union B.

OK, now let's stop for a little bit. We've computed the probability for each expression in the following fractional form. And we observed that for both fractions, the numerator is the same. So the numerator is a probability of A and B. And the denominator in the first case is probably of A, and the second case, probably of A union B.

Since we know that A is a subset of the event A union B, and by the monotonicity of probabilities, we know that the probability of A is hence no greater than a probability of A union B. Substituting this back into these expressions, we know that because they lie in the denominators, the first expression is indeed no smaller than the second expression. So our friend Alice was correct.

So throughout this problem, we never used the fact that the probability of a particular coin toss results, let's say, in a head is a certain number. Actually, this bias for the coin is irrelevant. Whether the coin is fair or unfair, this fact is always true.

So indeed, it does not depend on the probability of the coin. But if you're really curious what happens when the coin is fair, we can plug in the numbers. And here, we're assuming the coin is fair, which means probability of having a head is 1/2. Then, we'll see after going through the calculations that the first probability is 1/2, whereas the second probability is 1/3, which means, in this case, the [? dominance ?] actually is strict. So the first one is strictly greater than the second one, OK?

So this completes the first part of the problem. How do we generalize this into more general settings? There are multiple ways, but we'll go over one particular form. And to do so, we'll be defining three events somewhat more abstractly.

Let's say we have three events-- C, D, and E. Imagine any event, but all three events have to satisfy the following condition. First, event D will be a subset of E. And second, the intersection of C and D is equal to the intersection of C and E, OK? So this will be our choice events.

And let's see a particular example. Let's say you have a sample space here and some event E. Now, by the first condition, D will have to lie somewhere in E. For the second condition, we'll pick some event C such that this is true. And one way to do so is simply picking C that lies within both D and E. And you can see C intersection D will be C. And C intersection E will still be C. Hence, the second equality is true.

So if both equalities are true, we have the following relationship, that the probability of C conditional on D will be no smaller than the probability of C conditional on event E. And this will be the more general form of the inequality that we saw before.

So first of all, the way to prove this is in fact the same. We simply write out the value of this using the fractional form. And based on these two facts, we can arrive at this equation, which I shall now go over. But just to see why this form is more general, in fact, if we-- say we let C be the event A intersection B, D be the event A, and E be the event A and B where A and B are the events that we defined earlier.

We can verify that, indeed, these conditions are true, namely D is a subset of E. Because A is a subset of A union B, and C is a subset of both D and E. And hence, condition two is also true. And if that's the case, we will actually recover the result we got earlier for events A and B. And hence, this equation here is a more general form. So that's the end of the problem. See you next time.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013 Transcript – Recitation: Conditional Probability Example

Hi. Today we're going to do another fun problem that involves rolling two dice. So if you guys happen to frequent casinos, this problem might be really useful for you. I'm just kidding.

But in all seriousness, this problem is a good problem, because it's going to remind us how and when to use the discrete uniform law. Don't worry, I'll review what that says. And it's also going to exercise your understanding of conditional probability.

So quick recap. The discrete uniform law says that when your sample space is discrete, and when the outcomes in your sample space are equally likely, then to compute the probability of any event A, you can simply count the number of outcomes in A and divide it by the total number of possible outcomes. OK, so coming back to our problem. The problem statement tells us that we roll two fair six-sided die. And it also tells us that each one of the 36 possible outcomes is assumed to be equally likely.

So you know alarm bell should be going off in your head. Our sample space is clearly discrete. And it says explicitly that all outcomes are equally likely. So clearly, we can use the discrete uniform law. And again, this is helpful because it reduces a problem of computing probabilities to a problem of counting.

OK, and before we go any further, I just want to review what this graph is plotting. You've seen it a few times, but just to clarify, on one axis, we're plotting the outcome of the first die roll, and on the second axis, we're plotting the outcome of the second die roll. So if you got a 4 on your first die, and you get a 1 on your second die, that corresponds to this point over 4 and up 1.

OK, so part a asks us to find the probability that doubles our rolls. So let's use some shorthand. We're going to let D be the event that doubles are rolled. And we want to compute the probability of D.

I argue before we can use the discrete uniform law. So if we apply that, we just get the number of outcomes that comprise the event "doubles rolled" divided by 36, because there are 36 possible outcomes, which you can see just by counting the dots in this graph. Six possible outcomes for the first die, six possible outcomes for the second die. That's how you-- 6 times 6 is 36.

So I've been assuming this entire time that you know what doubles are. For those of you who don't know, doubles is essentially when that number on the first die matches the number on the second die. So this outcome here 1-1 is part of the event "doubles rolled." Similarly, 2-2, 3-3, 4-4, 5-5, and 6-6-- these six points comprise the event "doubles rolled."

So we can go ahead and put 6 over 36, which is equal to 1/6. So we're done with part a. We haven't seen any conditioning yet. The conditioning comes in part b.

So in part b we're still interested in the event D, in the event that "doubles are rolled." But now we want to compute this probability conditioned on the event that the sum of the results is less than or equal to 4. So I'm going to use this shorthand sum less than or equal to 4 to denote the event that the role results in the sum of 4 or smaller.

So there's two ways we're going to go about solving part b. Let's just jump right into the first way. The first way is applying the definition of conditional probability. So hopefully you remember that this is just probability of D intersect sum less than or equal to 4, divided by probability of sum less than or equal to 4.

Now, sum less than or equal to 4 and D intersect sum less than or equal to 4 are just two events. And so we can apply the discrete uniform law to calculate both the numerator and the denominator. So let's start with the denominator first because it seems a little bit easier.

So sum less than or equal to 4, let's figure this out. Well, 1-1 gives us a sum of 2, that's less than or equal to 4. 2-1 gives us 3. 3-1 gives us 4. 4-1 gives us 5, so we don't want to include this or this, or this point.

And you can sort of convince yourself that the next point we want to include is this one. That corresponds to 2-2, which is 4, so it makes sense that these guys should form the boundary, because all dots sort of up and to the right will have a bigger sum.

3-1 gives us 4. And 1-2 gives us 3. So these six points-- 1, 2, 3, 4, 5, 6-- are the outcomes that comprise the event sum less than or equal to 4.

So we can go ahead and write in the denominator, 6 over 36, because we just counted the outcomes in sum less than or equal to 4 and divided it by the number of outcomes in omega. Now, let's compute the numerator. D intersect sum less than or equal to 4. So we already found the blue check marks. Those correspond to sum less than or equal to 4.

Out of the points that have blue check marks, which one correspond to doubles? Well, they're actually already circled. It's just these two points. So we don't even need to circle those, so we get 2 over 36, using the discrete uniform law. And you see that these two 36s cancel each other. So you just get 2/6 or 1/3.

So that is one way of solving part b, but I want to take you, guys, through a different way, which I think is important, and that make sure you really understand what conditioning means. So another way that you can solve part b is to say, OK, we are now in the universe, we are in the conditional universe, where we know the sum of our results is 4 or smaller. And so that means our new sample space is really just this set of six points.

And one thing that it's worth noting is that conditioning never changes the relative frequencies or relative likelihoods of the different outcomes. So because all outcomes were equally likely in our original sample space omega, in the conditional worlds the outcomes are also equally likely. So using that argument, we could say that in our sort of blue conditional universe all of the

outcomes are equally likely. And therefore, we can apply a conditional version of the discrete uniform law.

So namely, to compute the probability of some event in that conditional world. So the conditional probability that "doubles are rolled", we need only count the number of outcomes in that event and divide it by the total number of outcomes.

So in the conditional world, there's only two outcomes that comprise the event "doubles rolled." These are the only two circles in the blue region, right? So applying the conditional version number law, we have two. And then we need to divide by the size of omega. So our conditional universe, we've already said, has six possible dots. So we just divide by 6, and you see that we get the same answer of 1/3.

And so again, we used two different strategies. I happen to prefer the second one, because it's slightly faster and it makes you think about what does conditioning really mean. Conditioning means you're now restricting your attention to a conditional universe. And given that you're in this conditional universe where the sum was less than or equal to 4, what is then the probability that doubles also happened?

OK, hopefully you, guys, are following. Let's move on to part c. So part c asks for the probability that at least one die roll is a 6. So I'm going to use the letter S to denote this, the probability that at least one die roll is a 6.

So let's go back to our picture and we'll use a green marker. So hopefully you agree that anything in this column corresponds to at least one 6. So this point, this point, this point, this point, this point, and this point your first die landed on a 6, so at least one 6 is satisfied. Similarly, if your second die has a 6, then we're also OK.

So I claim we want to look at these 11 points. Let me just check that, yeah, 6 plus 5-- 11. So using the discrete uniform law again, we get 11 divided by 36.

OK, last problem, we're almost done. So again, we're interested in the event S again, so the event that at least one die roll is a 6. But now we want to compute the probability of that event in the conditional world where the two dice land on different numbers.

So I'm going to call this probability of S. Let's see, I'm running out of letters. Let's for lack of a better letter, my name is Katie, so we'll just use a K. We want to compute the probability of S given K. And instead of using the definition of conditional probability, like we did back in part b, we're going to use the faster route.

So essentially, we're going to find the number of outcomes in the conditional world. And then we're also going to compute the number of outcomes that comprise S in the conditional world. So let's take a look at this. We are conditioning on the event that the two dice land on different numbers.

So hopefully you agree with me that every single dot that is not on the diagonal, so every single dot that doesn't correspond to doubles, is a dot that we care about. So our conditional universe of that the two dice land on "different numbers", that corresponds to these dots. And it corresponds to these dots. I don't want to get this one. OK, that's good.

So let's see, how many outcomes do we have in our conditional world? And I'm sorry I don't know why I didn't include this. This is absolutely included. I'm just testing to see if you, guys, are paying attention.

So we counted before that there are six dots on the diagonal, and we know that there are 36 dots total. So the number of dots, or outcomes to use the proper word, in our conditional world is 36 minus 6, or 30. So we get a 30 on the denominator.

And now we're sort of using a conditional version of our discrete uniform law, again. And the reason why we can do this is, as I argued before, that conditioning doesn't change the relative frequency of the outcomes. So in this conditional world, all of the outcomes are still equally likely, hence we can apply this law again.

So now we need to count the number of outcomes that are in the orange conditional world, but that also satisfy at least one die roll is a 6. So you can see-- 1-- we just need to count the green circles that are also in the orange. So that's 1, 2, 3, 4, 5, 6, 7, 8, 9, 10. So we get a 10, so our answer is 10 over 30, or 1/3.

So now we're done with this problem. As you see, hopefully, it wasn't too painful. And what are the important takeaways here for this problem?

Well, one is that whenever you have a discrete sample space, in which all of outcomes are equally likely, you should think about using the discrete uniform law, because this law lets you reduce the problem from computing probabilities to just counting outcomes within events. And the second takeaway is the way we thought about conditioning.

So we talked about one thing, which is that in your conditional world, when you condition, the relative likelihoods of the various outcomes don't change. So in our original universe, all of the outcomes were equally likely. So in our conditional universe, all of the outcomes are equally likely. And we saw it was much faster to apply a conditional version of the discrete uniform law. So that's it for today. And we'll do more problems next time.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013 Transcript – Lecture 2

The following content is provided under a Creative Commons license. Your support will help MIT OpenCourseWare continue to offer high quality educational resources for free. To make a donation or view additional materials from hundreds of MIT courses, visit MIT OpenCourseWare at ocw.mit.edu

JOHN TSISIKLIS: So here's the agenda for today. We're going to do a very quick review. And then we're going to introduce some very important concepts. The idea is that all information is-- Information is always partial. And the question is what do we do to probabilities if we have some partial information about the random experiments. We're going to introduce the important concept of conditional probability. And then we will see three very useful ways in which it is used. And these ways basically correspond to divide and conquer methods for breaking up problems into simpler pieces. And also one more fundamental tool which allows us to use conditional probabilities to do inference, that is, if we get a little bit of information about some phenomenon, what can we infer about the things that we have not seen?

So our quick review. In setting up a model of a random experiment, the first thing to do is to come up with a list of all the possible outcomes of the experiment. So that list is what we call the sample space. It's a set. And the elements of the sample space are all the possible outcomes. Those possible outcomes must be distinguishable from each other. They're mutually exclusive. Either one happens or the other happens, but not both. And they are collectively exhaustive, that is no matter what the outcome of the experiment is going to be an element of the sample space.

And then we discussed last time that there's also an element of art in how to choose your sample space, depending on how much detail you want to capture. This is usually the easy part. Then the more interesting part is to assign probabilities to our model, that is to make some statements about what we believe to be likely and what we believe to be unlikely. The way we do that is by assigning probabilities to subsets of the sample space. So as we have our sample space here, we may have a subset A. And we assign a number to that subset $P(A)$, which is the probability that this event happens. Or this is the probability that when we do the experiment and we get an outcome it's the probability that the outcome happens to fall inside that event.

We have certain rules that probabilities should satisfy. They're non-negative. The probability of the overall sample space is equal to one, which expresses the fact that we're are certain, no matter what, the outcome is going to be an element of the sample space. Well, if we set the top right so that it exhausts all possibilities, this should be the case.

And then there's another interesting property of probabilities that says that, if we have two events or two subsets that are disjoint, and we're interested in the probability, that one or the other happens, that is the outcome belongs to A or belongs to B. For disjoint events the total probability of these two, taken together, is just the sum of their individual probabilities. So probabilities behave like masses. The mass of the object consisting of A and B is the sum of the masses of these two objects. Or you can think of probabilities as areas. They have, again, the same property. The area of A together with B is the area of A plus the area B.

But as we discussed at the end of last lecture, it's useful to have in our hands a more general version of this additivity property, which says the following, if we take a sequence of sets-- A1, A2, A3, A4, and so on. And we put all of those sets together. It's an infinite sequence. And we ask for the probability that the outcome falls somewhere in this infinite union, that is we are asking for the probability that the outcome belongs to one of these sets, and assuming that the sets are disjoint, we can again find the probability for the overall set by adding up the probabilities of the individual sets.

So this is a nice and simple property. But it's a little more subtle than you might think. And let's see what's going on by considering the following example. We had an example last time where we take our sample space to be the unit square. And we said let's consider a probability law that says that the probability of a subset is just the area of that subset. So let's consider this probability law. OK.

Now the unit square is the set --let me just draw it this way-- the unit square is the union of one element set consisting all of the points. So the unit square is made up by the union of the various points inside the square. So union over all x's and y's. OK? So the square is made up out of all the points that this contains.

And now let's do a calculation. One is the probability of our overall sample space, which is the unit square. Now the unit square is the union of these things, which, according to our additivity axiom, is the sum of the probabilities of all of these one element sets. Now what is the probability of a one element set? What is the probability of this one element set? What's the probability that our outcome is exactly that particular point? Well, it's the area of that set, which is zero. So it's just the sum of zeros. And by any reasonable definition the sum of zeros is zero. So we just proved that one is equal to zero.

OK. Either probability theory is dead or there is some mistake in the derivation that I did. OK, the mistake is quite subtle and it comes at this step. We're sort of applied the additivity axiom by saying that the unit square is the union of all those sets. Can we really apply our additivity axiom. Here's the catch. The additivity axiom applies to the case where we have a sequence of disjoint events and we take their union. Is this a sequence of sets? Can you make up the whole unit square by taking a sequence of elements inside it and cover the whole unit square? Well if you try, if you start looking at the sequence of one element points, that sequence will never be able to exhaust the whole unit square.

So there's a deeper reason behind that. And the reason is that infinite sets are not all of the same size. The integers are an infinite set. And you can arrange the integers in a sequence. But the continuous set like the units square is a bigger set. It's so-called uncountable. It has more elements than any sequence could have. So this union here is not of this kind, where we would have a sequence of events. It's a different kind of union. It's a Union that involves a union of many, many more sets. So the countable additivity axiom does not apply in this case. Because, we're not dealing with a sequence of sets. And so this is the incorrect step.

So at some level you might think that this is puzzling and awfully confusing. On the other hand, if you think about areas of the way you're used to them from calculus, there's nothing mysterious

about it. Every point on the unit square has zero area. When you put all the points together, they make up something that has finite area. So there shouldn't be any mystery behind it.

Now, one interesting thing that this discussion tells us, especially the fact that the single elements set has zero area, is the following-- Individual points have zero probability. After you do the experiment and you observe the outcome, it's going to be an individual point. So what happened in that experiment is something that initially you thought had zero probability of occurring. So if you happen to get some particular numbers and you say, "Well, in the beginning, what did I think about those specific numbers? I thought they had zero probability. But yet those particular numbers did occur."

So one moral from this is that zero probability does not mean impossible. It just means extremely, extremely unlikely by itself. So zero probability things do happen. In such continuous models, actually zero probability outcomes are everything that happens. And the bumper sticker version of this is to always expect the unexpected. Yes?

AUDIENCE: [INAUDIBLE].

JOHN TSISIKLIS: Well, probability is supposed to be a real number. So it's either zero or it's a positive number. So you can think of the probability of things just close to that point and those probabilities are tiny and close to zero. So that's how we're going to interpret probabilities in continuous models. But this is two chapters ahead. Yeah?

AUDIENCE: How do we interpret probability of zero? If we can use models that way, then how about probability of one? That it it's extremely likely but not necessarily for certain?

JOHN TSISIKLIS: That's also the case. For example, if you ask in this continuous model, if you ask me for the probability that x, y , is different than the zero, zero this is the whole square, except for one point. So the area of this is going to be one. But this event is not entirely certain because the zero, zero outcome is also possible. So again, probability of one means essential certainty. But it still allows the possibility that the outcome might be outside that set. So these are some of the weird things that are happening when you have continuous models. And that's why we start to this class with discrete models, on which would be spending the next couple of weeks.

OK. So now once we have set up our probability model and we have a legitimate probability law that has these properties, then the rest is usually simple. Somebody asks you a question of calculating the probability of some event. While you were told something about the probability law, such as for example the probabilities are equal to areas, and then you just need to calculate. In these type of examples somebody would give you a set and you would have to calculate the area of that set. So the rest is just calculation and simple.

Alright, so now it's time to start with our main business for today. And the starting point is the following-- You know something about the world. And based on what you know when you set up a probability model and you write down probabilities for the different outcomes. Then something happens, and somebody tells you a little more about the world, gives you some new

information. This new information, in general, should change your beliefs about what happened or what may happen. So whenever we're given new information, some partial information about the outcome of the experiment, we should revise our beliefs. And conditional probabilities are just the probabilities that apply after the revision of our beliefs, when we're given some information.

So let's make this into a numerical example. So inside the sample space, this part of the sample space, let's say has probability $3/6$, this part has $2/6$, and that part has $1/6$. I guess that means that out here we have zero probability. So these were our initial beliefs about the outcome of the experiment. Suppose now that someone comes and tells you that event B occurred. So they don't tell you the full outcome with the experiment. But they just tell you that the outcome is known to lie inside this set B.

Well then, you should certainly change your beliefs in some way. And your new beliefs about what is likely to occur and what is not is going to be denoted by this notation. This is the conditional probability that the event A is going to occur, the probability that the outcome is going to fall inside the set A given that we are told and we're sure that the event lies inside the event B. Now once you're told that the outcome lies inside the event B, then our old sample space in some ways is irrelevant. We have then you sample space, which is just the set B. We are certain that the outcome is going to be inside B.

For example, what is this conditional probability? It should be one. Given that I told you that B occurred, you're certain that B occurred, so this has unit probability. So here we see an instance of revision of our beliefs. Initially, event B had the probability of $(2+1)/6$ -- that's $1/2$. Initially, we thought B had probability $1/2$. Once we're told that B occurred, the new probability of B is equal to one. OK.

How do we revise the probability that A occurs? So we are going to have the outcome of the experiment. We know that it's inside B. So we will either get something here, and A does not occur. Or something inside here, and A does occur. What's the likelihood that, given that we're inside B, the outcome is inside here? Here's how we're going to think about. This part of this set B, in which A also occurs, in our initial model was twice as likely as that part of B. So outcomes inside here collectively were twice as likely as outcomes out there.

So we're going to keep the same proportions and say, that given that we are inside the set B, we still want outcomes inside here to be twice as likely outcomes there. So the proportion of the probabilities should be two versus one. And these probabilities should add up to one because together they make the conditional probability of B. So the conditional probabilities should be $2/3$ probability of being here and $1/3$ probability of being there. That's how we revise our probabilities. That's a reasonable, intuitively reasonable, way of doing this revision. Let's translate what we did into a definition.

The definition says the following, that the conditional probability of A given that B occurred is calculated as follows. We look at the total probability of B. And out of that probability that was inside here, what fraction of that probability is assigned to points for which the event A also occurs? Does it give us the same numbers as we got with this heuristic argument? Well in this

example, probability of A intersection B is 2/6, divided by total probability of B, which is 3/6, and so it's 2/3, which agrees with this answer that's we got before. So the former indeed matches what we were trying to do.

One little technical detail. If the event B has zero probability, and then here we have a ratio that doesn't make sense. So in this case, we say that conditional probabilities are not defined.

Now you can take this definition and unravel it and write it in this form. The probability of A intersection B is the probability of B times the conditional probability. So this is just consequence of the definition but it has a nice interpretation. Think of probabilities as frequencies. If I do the experiment over and over, what fraction of the time is it going to be the case that both A and B occur? Well, there's going to be a certain fraction of the time at which B occurs. And out of those times when B occurs, there's going to be a further fraction of the experiments in which A also occurs.

So interpret the conditional probability as follows. You only look at those experiments at which B happens to occur. And look at what fraction of those experiments where B already occurred, event A also occurs. And there's a symmetrical version of this equality. There's symmetry between the events B and A. So you also have this relation that goes the other way.

OK, so what do we use these conditional probabilities for? First, one comment. Conditional probabilities are just like ordinary probabilities. They're the new probabilities that apply in a new universe where event B is known to have occurred. So we had an original probability model. We are told that B occurs. We revise our model. Our new model should still be legitimate probability model. So it should satisfy all sorts of properties that ordinary probabilities do satisfy.

So for example, if A and B are disjoint events, then we know that the probability of A union B is equal to the probability of A plus probability of B. And now if I tell you that a certain event C occurred, we're placed in a new universe where event C occurred. We have new probabilities for that universe. These are the conditional probabilities. And conditional probabilities also satisfy this kind of property. So this is just our usual additivity axiom but the applied in a new model, in which we were told that event C occurred. So conditional probabilities do not taste or smell any different than ordinary probabilities do. Conditional probabilities, given a specific event B, just form a probability law on our sample space. It's a different probability law but it's still a probability law that has all of the desired properties.

OK, so where do conditional probabilities come up? They do come up in quizzes and they do come up in silly problems. So let's start with this. We have this example from last time. Two rolls of a die, all possible pairs of roles are equally likely, so every element in this square has probability of 1/16. So all elements are equally likely. That's our original model. Then somebody comes and tells us that the minimum of the two rolls is equal to zero. What's that event? The minimum equal to zero can happen in many ways, if we get two zeros or if we get a zero and-- sorry, if we get two two's, or get a two and something larger. And so the is our new event B. The red event is the event B.

And now we want to calculate probabilities inside this new universe. For example, you may be interested in the question, questions about the maximum of the two rolls. In the new universe, what's the probability that the maximum is equal to one? The maximum being equal to one is this black event. And given that we're told that B occurred, this black events cannot happen. So this probability is equal to zero. How about the maximum being equal to two, given that event B? OK, we can use the definition here. It's going to be the probability that the maximum is equal to two and B occurs divided by the probability of B. The probability that the maximum is equal to two.

OK, what's the event that the maximum is equal to two? Let's draw it. This is going to be the blue event. The maximum is equal to two if we get any of those blue points. So the intersection of the two events is the intersection of the red event and the blue event. There's only one point in their intersection. So the probability of that intersection happening is $1/16$. That's the numerator. How about the denominator? The event B consists of five elements, each one of which had probability of $1/16$. So that's $5/16$. And so the answer is $1/5$.

Could we have gotten this answer in a faster way? Yes. Here's how it goes. We're trying to find the conditional probability that we get this point, given that B occurred. B consist of five elements. All of those five elements were equally likely when we started, so they remain equally likely afterwards. Because when we define conditional probabilities, we keep the same proportions inside the set. So the five red elements were equally likely. They remain equally likely in the conditional world. So conditional event B having happened, each one of these five elements has the same probability. So the probability that we actually get this point is going to be $1/5$. And so that's the shortcut.

More generally, whenever you have a uniform distribution on your initial sample space, when you condition on an event, your new distribution is still going to be uniform, but on the smaller events of that we considered. So we started with a uniform distribution on the big square and we ended up with a uniform distribution just on the red point.

Now besides silly problems, however, conditional probabilities show up in real and interesting situations. And this example is going to give you some idea of how that happens. OK. Actually, in this example, instead of starting with a probability model in terms of regular probabilities, I'm actually going to define the model in terms of conditional probabilities. And we'll see how this is done. So here's the story. There may be an airplane flying up in the sky, in a particular sector of the sky that you're watching. Sometimes there is one sometimes there isn't. And from experience you know that when you look up, there's five percent probability that the plane is flying above there and 95% probability that there's no plane up there.

So event A is the event that the plane is flying out there. Now you bought this wonderful radar that's looks up. And you're told in the manufacturer's specs that, if there is a plane out there, your radar is going to register something, a blip on the screen with probability 99%. And it will not register anything with probability one percent. So this particular part of the picture is a self-contained probability model of what your radar does in a world where a plane is out there. So I'm telling you that the plane is out there.

So we're now dealing with conditional probabilities because I gave you some particular information. Given this information that the plane is out there, that's how your radar is going to behave with probability 99% is going to detect it, with probability one percent is going to miss it. So this piece of the picture is a self-contained probability model. The probabilities add up to one. But it's a piece of a larger model.

Similarly, there's the other possibility. Maybe a plane is not up there and the manufacturer specs tell you something about false alarms. A false alarm is the situation where the plane is not there, but for some reason your radar picked up some noise or whatever and shows a blip on the screen. And suppose that this happens with probability ten percent. Whereas with probability 90% your radar gives the correct answer.

So this is sort of a model of what's going to happen with respect to both the plane -- we're given probabilities about this -- and we're given probabilities about how the radar behaves. So here I have indirectly specified the probability law in our model by starting with conditional probabilities as opposed to starting with ordinary probabilities. Can we derive ordinary probabilities starting from the conditional number ones? Yeah, we certainly can.

Let's look at this event, A intersection B, which is the event up here, that there is a plane and our radar picks it up. How can we calculate this probability? Well we use the definition of conditional probabilities and this is the probability of A times the conditional probability of B given A. So it's 0.05 times 0.99. And the answer, in case you care-- It's 0.0495. OK. So we can calculate the probabilities of final outcomes, which are the leaves of the tree, by using the probabilities that we have along the branches of the tree. So essentially, what we ended up doing was to multiply the probability of this branch times the probability of that branch.

Now, how about the answer to this question. What is the probability that our radar is going to register something? OK, this is an event that can happen in multiple ways. It's the event that consists of this outcome. There is a plane and the radar registers something together with this outcome, there is no plane but the radar still registers something.

So to find the probability of this event, we need the individual probabilities of the two outcomes. For the first outcome, we already calculated it. For the second outcome, the probability that this happens is going to be this probability 95% times 0.10, which is the conditional probability for taking this branch, given that there was no plane out there. So we just add the numbers. 0.05 times 0.99 plus 0.95 times 0.1 and the final answer is 0.1445. OK.

And now here's the interesting question. Given that your radar recorded something, how likely is it that there is an airplane up there? Your radar registering something -- that can be caused by two things. Either there's a plane there, and your radar did its job. Or there was nothing, but your radar fired a false alarm. What's the probability that this is the case as opposed to that being the case? OK. The intuitive shortcut would be that it should be the probability-- you look at their relative odds of these two elements and you use them to find out how much more likely it is to be there as opposed to being there.

But instead of doing this, let's just write down the definition and just use it. It's the probability of A and B happening, divided by the probability of B. This is just our definition of conditional probabilities. Now we have already found the numerator. We have already calculated the denominator. So we take the ratio of these two numbers and we find the final answer -- which is 0.34. OK.

There's this slightly curious thing that's happened in this example. Doesn't this number feel a little too low? My radar -- So this is a conditional probability, given that my radar said there is something out there, that there is indeed something there. So it's sort of the probability that our radar gave the correct answer. Now, the specs of our radar we're pretty good. In this situation, it gives you the correct answer 99% of the time. In this situation, it gives you the correct answer 90% of the time. So you would think that your radar there is really reliable.

But yet here the radar recorded something, but the chance that the answer that you get out of this is the right one, given that it recorded something, the chance that there is an airplane out there is only 30%. So you cannot really rely on the measurements from your radar, even though the specs of the radar were really good. What's the reason for this? Well, the reason is that false alarms are pretty common.

Most of the time there's nothing. And there's a ten percent probability of false alarms. So there's roughly a ten percent probability that in any given experiment, you have a false alarm. And there is about the five percent probability that something out there and your radar gets it. So when your radar records something, it's actually more likely to be a false alarm rather than being an actual airplane. This has probability ten percent roughly. This has probability roughly five percent

So conditional probabilities are sometimes counter-intuitive in terms of the answers that they get. And you can make similar stories about doctors interpreting the results of tests. So you tested positive for a certain disease. Does it mean that you have the disease necessarily? Well if that disease has been eradicated from the face of the earth, testing positive doesn't mean that you have the disease, even if the test was designed to be a pretty good one. So unfortunately, doctors do get it wrong also sometimes. And the reasoning that comes in such situations is pretty subtle.

Now for the rest of the lecture, what we're going to do is to take this example where we did three things and abstract them. These three trivial calculations that's we just did are three very important, very basic tools that you use to solve more general probability problems. So what's the first one? We find the probability of a composite event, two things happening, by multiplying probabilities and conditional probabilities. More general version of this, look at any situation, maybe involving lots and lots of events.

So here's a story that event A may happen or may not happen. Given that A occurred, it's possible that B happens or that B does not happen. Given that B also happens, it's possible that the event C also happens or that event C does not happen. And somebody specifies for you a model by giving you all these conditional probabilities along the way. Notice what we move along the branches as the tree progresses. Any point in the tree corresponds to certain events having happened.

And then, given that this has happened, we specify conditional probabilities. Given that this has happened, how likely is it for that C also occurs? Given a model of this kind, how do we find the probability or for this event? The answer is extremely simple. All that you do is move along with the tree and multiply conditional probabilities along the way. So in terms of frequencies, how often do all three things happen, A, B, and C? You first see how often does A occur. Out of the times that A occurs, how often does B occur? And out of the times where both A and B have occurred, how often does C occur? And you can just multiply those three frequencies with each other.

What is the formal proof of this? Well, the only thing we have in our hands is the definition of conditional probabilities. So let's just use this. And-- OK. Now, the definition of conditional probabilities tells us that the probability of two things is the probability of one of them times a conditional probability. Unfortunately, here we have the probability of three things. What can I do? I can put a parenthesis in here and think of this as the probability of this and that and apply our definition of conditional probabilities here. The probability of two things happening is the probability that the first happens times the conditional probability that the second happens, given A and B, given that the first one happened.

So this is just the definition of the conditional probability of an event, given another event. That other event is a composite one, but that's not an issue. It's just an event. And then we use the definition of conditional probabilities once more to break this apart and make it $P(A)$, $P(B \text{ given } A)$ and then finally, the last term. OK.

So this proves the formula that I have up there on the slides. And if you wish to calculate any other probability in this diagram. For example, if you want to calculate this probability, you would still multiply the conditional probabilities along the different branches of the tree. In particular, here in this branch, you would have the conditional probability of C complement, given A intersection B complement, and so on. So you write down probabilities along all those tree branches and just multiply them as you go.

So this was the first skill that we are covering. What was the second one? What we did was to calculate the total probability of a certain event B that consisted of-- was made up from different possibilities, which corresponded to different scenarios. So we wanted to calculate the probability of this event B that consisted of those two elements.

Let's generalize. So we have our big model. And this sample space is partitioned in a number of sets. In our radar example, we had a partition in two sets. Either a plane is there, or a plane is not there. Since we're trying to generalize, now I'm going to give you a picture for the case of three possibilities or three possible scenarios. So whatever happens in the world, there are three possible scenarios, A_1 , A_2 , A_3 . So think of these as there's nothing in the air, there's an airplane in the air, or there's a flock of geese flying in the air. So there's three possible scenarios.

And then there's a certain event B of interest, such as a radar records something or doesn't record something. We specify this model by giving probabilities for the A_i 's-- That's the probability of the different scenarios. And somebody also gives us the probabilities that this event B is going to occur, given that the A_i -th scenario has occurred. Think of the A_i 's as scenarios.

And we want to calculate the overall probability of the event B. What's happening in this example? Perhaps, instead of this picture, it's easier to visualize if I go back to the picture I was using before. We have three possible scenarios, A1, A2, A3. And under each scenario, B may happen or B may not happen. And so on. So here we have A2 intersection B. And here we have A3 intersection B. In the previous slide, we found how to calculate the probability of any event of this kind, which is done by multiplying probabilities here and conditional probabilities there.

Now we are asked to calculate the total probability of the event B. The event B can happen in three possible ways. It can happen here. It can happen there. And it can happen here. So this is our event B. It consists of three elements. To calculate the total probability of our event B, all we need to do is to add these three probabilities. So B is an event that consists of these three elements. There are three ways that B can happen. Either B happens together with A1, or B happens together with A2, or B happens together with A3.

So we need to add the probabilities of these three contingencies. For each one of those contingencies, we can calculate its probability by using the multiplication rule. So the probability of A1 and B happening is this-- It's the probability of A1 and then B happening given that A1 happens. The probability of this contingency is found by taking the probability that A2 happens times the conditional probability of A2, given that B happened. And similarly for the third one. So this is the general rule that we have here. The rule is written for the case of three scenarios. But obviously, it has a generalization for the case of four or five or more scenarios. It gives you a way of breaking up the calculation of an event that can happen in multiple ways by considering individual probabilities for the different ways that the event can happen.

OK. So-- Yes?

AUDIENCE: Does this have to change for infinite sample space?

JOHN TSISIKLIS: No. This is true whether your sample space is infinite or finite. What I'm using in this argument that we have a partition into just three scenarios, three events. So it's a partition to a finite number of events. It's also true if it's a partition into an infinite sequence of events. But that's, I think, one of the theoretical problems at the end of the chapter. You probably may not need it for now.

OK, going back to the story here. There are three possible scenarios about what could happen in the world that are captured here. Event, under each scenario, event B may or may not happen. And so these probabilities tell us the likelihoods of the different scenarios. These conditional probabilities tell us how likely is it for B to happen under one scenario, or the other scenario, or the other scenario.

The overall probability of B is found by taking some combination of the probabilities of B in the different possible worlds, in the different possible scenarios. Under some scenario, B may be very likely. Under another scenario, it may be very unlikely. We take all of these into account and weigh them according to the likelihood of the scenarios. Now notice that since A1, A2, and A3 form a partition, these three probabilities have what property? Add to what? They add to one. So it's the probability of this branch, plus this branch, plus this branch. So what we have

here is a weighted average of the probabilities of the B's into the different worlds, or in the different scenarios.

Special case. Suppose the three scenarios are equally likely. So P of A_1 equals $1/3$, equals to P of A_2 , P of A_3 . what are we saying here? In that case of equally likely scenarios, the probability of B is the average of the probabilities of B in the three different words, or in the three different scenarios. OK.

So to finally, the last step. If we go back again two slides, the last thing that we did was to calculate a conditional probability of this kind, probability of A given B , which is a probability associated essentially with an inference problem. Given that our radar recorded something, how likely is it that the plane was up there? So we're trying to infer whether a plane was up there or not, based on the information that we've got.

So let's generalize once more. And we're just going to rewrite what we did in that example, but in terms of general symbols instead of the specific numbers. So once more, the model that we have involves probabilities of the different scenarios. These we call them prior probabilities. They're are our initial beliefs about how likely each scenario is to occur. We also have a model of our measuring device that tells us under that scenario how likely is it that our radar will register something or not. So we're given again these conditional probabilities. We're given the conditional probabilities for these branches.

Then we are told that event B occurred. And on the basis of this new information, we want to form some new beliefs about the relative likelihood of the different scenarios. Going back again to our radar example, an airplane was present with probability 5%. Given that the radar recorded something, we're going to change our beliefs. Now, a plane is present with probability 34%. The radar, since we saw something, we are going to revise our beliefs as to whether the plane is out there or is not there.

And so what we need to do is to calculate the conditional probabilities of the different scenarios, given the information that we got. So initially, we have these probabilities for the different scenarios. Once we get the information, we update them and we calculate our revised probabilities or conditional probabilities given the observation that we made. OK. So what do we do? We just use the definition of conditional probabilities twice. By definition the conditional probability is the probability of two things happening divided by the probability of the conditioning event.

Now, I'm using the definition of conditional probabilities once more, or rather I use the multiplication rule. The probability of two things happening is the probability of the first and the second. So these are things that are given to us. They're the probabilities of the different scenarios. And it's the model of our measuring device, which we assume to be available. And how about the denominator? This is total probability of the event B . But we just found that's it's easy to calculate using the formula in the previous slide. To find the overall probability of event B occurring, we look at the probabilities of B occurring under the different scenario and weigh them according to the probabilities of all the scenarios.

So in the end, we have a formula for the conditional probability, A's given B, based on the data of the problem, which were probabilities of the different scenarios and conditional probabilities of B, given the A's. So what this calculation does is, basically, it reverses the order of conditioning. We are given conditional probabilities of these kind, where it's B given A and we produce new conditional probabilities, where things go the other way.

So schematically, what's happening here is that we have model of cause and effect and-- So a scenario occurs and that may cause B to happen or may not cause it to happen. So this is a cause/effect model. And it's modeled using probabilities, such as probability of B given A_i . And what we want to do is inference where we are told that B occurs, and we want to infer whether A_i also occurred or not. And the appropriate probabilities for that are the conditional probabilities that A occurred, given that B occurred.

So we're starting with a causal model of our situation. It models from a given cause how likely is a certain effect to be observed. And then we do inference, which answers the question, given that the effect was observed, how likely is it that the world was in this particular situation or state or scenario.

So the name of the Bayes rule comes from Thomas Bayes, a British theologian back in the 1700s. It actually-- This calculation addresses a basic problem, a basic philosophical problem, how one can learn from experience or from experimental data and some systematic way. So the British at that time were preoccupied with this type of question. Is there a basic theory that about how we can incorporate new knowledge to previous knowledge. And this calculation made an argument that, yes, it is possible to do that in a systematic way. So the philosophical underpinnings of this have a very long history and a lot of discussion around them. But for our purposes, it's just an extremely useful tool. And it's the foundation of almost everything that gets done when you try to do inference based on partial observations. Very well. Till next time.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013

Transcript – Recitation: The Monty Hall Problem

Hi. In the session, we'll be solving the Monty Hall problem. And this problem is based on an old game show that was called "Let's Make a Deal." And the host of this game show, his name was Monty Hall, which is why this problem is now known as the Monty Hall problem.

And this problem is actually pretty well-known, because there was some disagreement at the time over what the right answer to this problem should be. Even some really smart people didn't agree on what the right answer should be. And part of what might explain that disagreement is that they probably were considering slightly different variations of the problem, because as in all probability problems, the assumptions that you're working with are very important, because otherwise you may be solving an actually different problem. And so what we'll do first is really lay out concretely what all the assumptions are, what the rules of the game are. And then we'll go through the methodology to solve for the actual answer.

So the game is actually relatively simple. So you're on a game show and you're presented with three doors. These doors are closed. And behind one of these doors is a prize, let's say, a car. And behind the other two doors, there's nothing. You don't know which one it is.

And the rules of the game are that, first, you get to choose any one of these three. So you pick one of the doors that you want. They don't show you what's behind that door, but your friend, who actually knows which door has the prize behind it, will look at the remaining doors.

So let's, just for example, let's say you chose door one. Your friend will look at the other two doors and open one of them. And you will make sure that the one that he opens is empty. That is the prize not behind that one.

And at this point, one of the doors is open and its empty, you have your original door plus another unopened door. And you're given an option-- you could either stay with your initial choice or you can switch to the other unopened door. And whichever one is your final choice, they will open that door. And if there's a price behind it, you win, and if there not, then you don't win.

So the question that we're trying to answer is what is the better strategy here? Is the better strategy to stay with your initial choice or is it better to switch to the other unopened door?

OK, so it turns out that the specific rules here actually are very important. Specifically, the rule about how your friend chooses to open doors. And the fact that he will always open one of the two other door that you haven't picked and he will make sure that that door doesn't have a prize behind it. And let's see how that actually plays out in this problem.

So the simplest way, I think, of thinking about this problem is just to think about under what circumstances does staying with your initial choice win? So if you think about it, the only way

that you can win by staying with your initial choice is if your initial choice happened to be the door that has a prize behind it. And because you're sticking with the initial choice, you can actually kind of forget about the rest of the game, about opening of the other door and about switching.

It's as if you're playing a simpler game, which is just you have three doors, one of them has a prize behind it, and you choose one of them. And if you guessed right, then you win. If you didn't, then you don't win. And because the another important assumption is that the prize has an equal probability of being behind any one of three doors so one third, one third, one third.

Because of that, then if you stay with your first choice, you win only if your first choice happened to be the right one. And that is the case with probably one third. So with that simple argument you can convince yourself that the probability of winning, given the strategy of staying with your first choice, is one third.

Now, let's think about the other strategy, which is to switch. So under what circumstances does switching win for you? Well, if your first choice happened to be the right door, then switching away from that door will always lose. But let's say, that happens with probably one third. But the rest of the time with probably 2/3, your first choice would be wrong.

So let's give an example here. Let's say, the prize, which I'll denote by happy face, is behind door two. And your first choice was door one. So your first choice was wrong.

Now, your friend can open door two, because door two has the prize behind it. He also doesn't open the door that you initially picked. So he has to open door three. So door three is open, and now you have an option of sticking with your first choice-- door one-- or switching to door two. So in this case, it's obvious to see that switching wins for you.

And now, if instead, you picked door one first, and the prize was behind door three, again, you are wrong. And again, your friend is forced to open door two. And switching, again, wins for you.

And so if you think about it, switching will win for you, as long as your initial pick was wrong. If your initial pick was wrong, then the prize is behind one of the doors. Your friend has to open one of the doors, but he can't open the door that has the prize behind it. So he has to open the other bad door, leaving the good door with the prize behind it, as the one that you can switch to.

And so by switching you will win in this scenario. And what is the probability of that happening? Well, that happens if your initial pick was wrong, which happens with probably 2/3.

So the final answer then, it's pretty simple, the probability of winning if you stay is one third, and the probability of winning if you switch is 2/3. And so maybe counterintuitively the result is that it's actually better for you, twice as good for you, to switch rather than stay. And so that was the argument, the kind of simple argument.

We can also be more methodical about this and actually list out all of the possible outcomes. Because it's relatively small problem-- there's only three doors-- we can actually just list out all the possible outcomes. So for example, if you chose door one first, and the prize was behind door one, your friend has a choice. He can open door two or door three, because they're both empty. And then in that case, if you stay, you win, you picked the door correctly. And if you switch to two or three, then you lose.

But if you chose door one, the prize is behind door two, then your friend has to open door three, he is forced to do that, then staying will lose but switching would win. And so on for the other cases. And so again, this is just an exhaustive list of all the possible outcomes, from which you can see that, in fact, staying wins, only if your first choice was correct. And switching wins in all the other cases. And so one third of the time, staying would win, 2/3 of the time switching would win.

OK, so now, we have the answer. Let's try to figure out and convince ourselves that it is actually right, because you might think before going through this process that maybe it doesn't matter whether you stay or you switch, they both have the same probability of winning, or maybe even staying is better. So why is staying worse and switching better?

Well, the first argument really is something that we've already talked about. By staying, you're essentially banking on your first choice being correct, which is a relatively poor bet, because you have only one in three chance of being right. But by switching, you're actually banking on your first choice being wrong, which is a relatively better bet, because you're more likely to be wrong than right in your first choice, because you're just picking blindly. OK, so that is one intuitive explanation for why switching is better.

Another slightly different way to think about it is that instead of picking single doors, you're actually picking groups of doors. So let's say that your first pick was door one. Then you're actually really deciding between door one or doors two and three combined. So why is that?

It's because by staying with door one, you're staying with door one. But by switching, you're actually getting two doors for the price of one, because you know that your friend will reveal one of these to be empty, and the other one will stay closed. But switching really kind of buys you both of these. And so because it buys you two opportunities to win, you get 2/3 chance of winning, versus a one third chance.

Another way of thinking about this is to increase the scale of the problem, and maybe that will help visualize the counterintuitive answer. So instead of having three doors, imagine that you have 1,000 doors that are closed. And again, one prize is behind one of the doors.

And the rules are similar-- you pick one door first, and then your friend will open 998 other doors. And these doors are guaranteed to be empty. And now you're left with your initial door plus one other door that is unopened.

So now the question is should you stay with your first choice or switch to your other choice? And it should be more intuitively obvious now that the better decision would be to switch,

because you're overwhelmingly more likely to have picked incorrectly for your first pick. You have only 1 in 1,000 chance of getting it right. So that is kind of just taking this to a bigger extreme and really driving home the intuition.

OK, so what we've really discovered is that the fact that the rules of the game are that your friend has to open one of the other two doors and cannot reveal the prize plays a big role in this problem. And that is an important assumption.

OK, so now let's think about a slightly different variation now. So a different strategy. Instead of just always staying or always switching, we have a specific other strategy, which is that you will choose door one first and then, depending on what your friend does, you will act accordingly. So if your friend opens door two, you will not switch. And if your friend opens door three, you will switch. So let's draw out exactly what happens here.

So you have door one that you've chosen. And the prize can be behind doors one, two, or three. And again, it's equally likely. So the probabilities of these branches are one third, one third, and one third. And now given that, your friend in this scenario has a choice between opening doors two or three. And so because of doors, you chose one, the prize actually is behind one, and so two and three are both empty, so he can choose whichever one he wants to open.

And the problem actually hasn't specified how your friend actually decides between this. So we'll leave it in general. So we'll say that the probability p , your friend will open door two, in this case. And with the remaining probability 1 minus p , he will open door three.

What about in this case? Well, you chose door one. The prize is actually behind door two. So following the rules of the game, your friend is forced to open door three. So this happens with probability 1. And similarly, if the prize is behind door three, your friend is forced to open door two, which, again, happens with probably 1.

So now let's see how this strategy works. When do you win? You win when, according to the strategy, your final choice is the right door. So according to the strategy, in this case, your friend opened door two. And according to your strategy, if door two is open, you don't switch. So you stay with your first choice of one. And that happens to be the right one, so you win in this case.

But what about here? Your friend opened door three, and by your strategy, you do switch, which is the wrong choice here, so you lose. Here, you switch, because you open door three, and you switch to the right door, so that wins. And this one, you don't switch, and you lose.

All right, so what is the final probability of winning? And the final probability of winning is the probability of getting to these two outcomes, which happens with probability one third times p plus one third times 1. So one third. So the final answer is one third p plus one third.

And notice now that the answer isn't just a number. Like in this case, the answer was one third and $2/3$. And it didn't actually matter how your friend chose between these two doors when he had a choice. But in this case, it actually doesn't matter, because p stays in the answer.

But one thing that we can do is we can compare this with these strategies. So what we see is that, well p is a probability, so it has to be between 0 and 1. So this probability winning for this strategy is somewhere between one third times 0 plus one third, which is one third. And one third times 1 plus one third, which is $2/3$. So the strategy is somewhere between $2/3$ and one third.

So what we see is that no matter what, this strategy is at least as good as staying all the time, because that was only one third. And no matter what it can't be any better than switching, which was $2/3$. So you can also come up with lots of other different strategies and see what the probabilities of winning are in that case.

OK, so what have we learned in this problem? What are the key takeaways? One important takeaway is that it's important to really understand a problem and arrive at a concrete and precise set of assumptions. So really have a precise problem that you're solving.

And another important takeaway is to think about your final answer, make sure that that actually makes sense to you, make sure that you can justify it somehow intuitively. In that case, you can actually convince yourself that your answer is actually correct, because sometimes go through a lot of formulas, and sometimes your formula may have an error in there somewhere. But you could take the final answer and ask yourself does this actually makes sense intuitively? That's often a very good check and sometimes you can catch errors in your calculations that way. OK so we'll see next time.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

LECTURE 2

- **Readings:** Sections 1.3-1.4

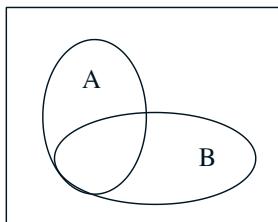
Lecture outline

- Review
- Conditional probability
- Three **important** tools:
 - Multiplication rule
 - Total probability theorem
 - Bayes' rule

Review of probability models

- **Sample space Ω**
 - Mutually exclusive
 - Collectively exhaustive
 - Right granularity
- **Event:** Subset of the sample space
- Allocation of probabilities to events
 1. $P(A) \geq 0$
 2. $P(\Omega) = 1$
 3. If $A \cap B = \emptyset$,
then $P(A \cup B) = P(A) + P(B)$
- 3'. If A_1, A_2, \dots are disjoint events, then:
 $P(A_1 \cup A_2 \cup \dots) = P(A_1) + P(A_2) + \dots$
- Problem solving:
 - Specify sample space
 - Define probability law
 - Identify event of interest
 - Calculate...

Conditional probability

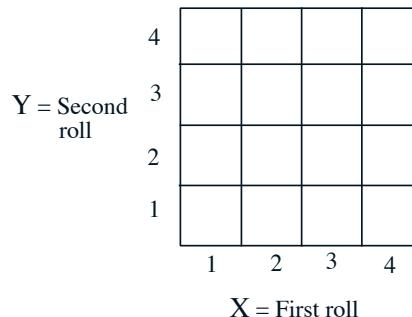


- $P(A | B) =$ probability of A , given that B occurred
 - B is our new universe
- **Definition:** Assuming $P(B) \neq 0$,

$$P(A | B) = \frac{P(A \cap B)}{P(B)}$$

$P(A | B)$ undefined if $P(B) = 0$

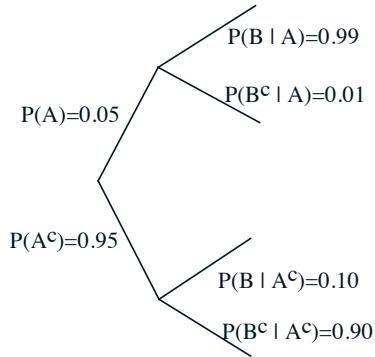
Die roll example



- Let B be the event: $\min(X, Y) = 2$
- Let $M = \max(X, Y)$
- $P(M = 1 | B) =$
- $P(M = 2 | B) =$

Models based on conditional probabilities

- Event A : Airplane is flying above
- Event B : Something registers on radar screen



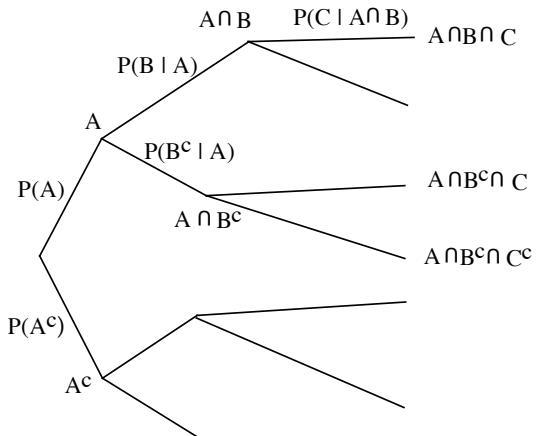
$$P(A \cap B) =$$

$$P(B) =$$

$$P(A | B) =$$

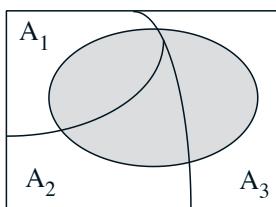
Multiplication rule

$$P(A \cap B \cap C) = P(A) \cdot P(B | A) \cdot P(C | A \cap B)$$



Total probability theorem

- Divide and conquer
- Partition of sample space into A_1, A_2, A_3
- Have $P(B | A_i)$, for every i

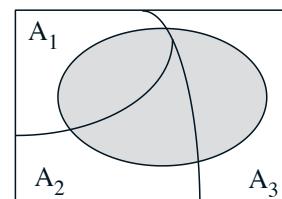


- One way of computing $P(B)$:

$$\begin{aligned} P(B) &= P(A_1)P(B | A_1) \\ &\quad + P(A_2)P(B | A_2) \\ &\quad + P(A_3)P(B | A_3) \end{aligned}$$

Bayes' rule

- "Prior" probabilities $P(A_i)$
 - initial "beliefs"
- We know $P(B | A_i)$ for each i
- Wish to compute $P(A_i | B)$
 - revise "beliefs", given that B occurred



$$\begin{aligned} P(A_i | B) &= \frac{P(A_i \cap B)}{P(B)} \\ &= \frac{P(A_i)P(B | A_i)}{P(B)} \\ &= \frac{P(A_i)P(B | A_i)}{\sum_j P(A_j)P(B | A_j)} \end{aligned}$$

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Recitation 2
September 14, 2010

1. Problem 1.15, page 56-57 in the text.

A coin is tossed twice. Alice claims that the event of two heads is at least as likely if we know that the first toss is a head than if we know that at least one of the tosses is a head. Is she right? Does it make a difference if the coin is fair or unfair? How can we generalize Alice's reasoning?

2. Problem 1.14, page 56 in the text.

We roll two fair 6-sided dice. Each one of the 36 possible outcomes is assumed to be equally likely.

- Find the probability that doubles are rolled.
- Given that the roll results in a sum of 4 or less, find the conditional probability that doubles are rolled.
- Find the probability that at least one die roll is a 6.
- Given that the two dice land on different numbers, find the conditional probability that at least one die roll is a 6.

3. Example 1.13, page 29, and Example 1.17, page 33, in the text.

You enter a chess tournament where your probability of winning a game is 0.3 against half of the players (call them type 1), 0.4 against a quarter of the players (call them type 2), and 0.5 against the remaining quarter of the players (call them type 3). You play a game against a randomly chosen opponent.

- What is the probability of winning?
- Suppose that you win. What is the probability that you had an opponent of type 1?

4. Example 1.12, page 27 in the text.

The Monty Hall Problem. This is a much discussed puzzle, based on an old American game show. You are told that a prize is equally likely to be found behind any one of three closed doors in front of you. You point to one of the doors. A friend opens for you one of the remaining two doors, after making sure that the prize is not behind it. At this point, you can stick to your initial choice, or switch to the other unopened door. You win the prize if it lies behind your final choice of a door. Consider the following strategies:

- Stick to your initial choice.
- Switch to the other unopened door.
- You first point to door 1. If door 2 is opened, you do not switch. If door 3 is opened, you switch.

Which is the best strategy?

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Recitation 2: Solutions
September 14, 2010

- Let A be the event that the first toss is a head and let B be the event that the second toss is a head. We must compare the conditional probabilities $\mathbf{P}(A \cap B|A)$ and $\mathbf{P}(A \cap B|A \cup B)$. We have

$$\mathbf{P}(A \cap B|A) = \frac{\mathbf{P}((A \cap B) \cap A)}{\mathbf{P}(A)} = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(A)},$$

and

$$\mathbf{P}(A \cap B|A \cup B) = \frac{\mathbf{P}((A \cap B) \cap (A \cup B))}{\mathbf{P}(A \cup B)} = \frac{A \cap B}{A \cup B}.$$

Since $\mathbf{P}(A \cup B) \geq \mathbf{P}(A)$, the first conditional probability above is at least as large, so Alice is right, regardless of whether the coin is fair or not. In the case where the coin is fair, that is, if all four outcomes HH, HT, TH, TT are equally likely, we have

$$\frac{\mathbf{P}(A \cap B)}{\mathbf{P}(A)} = \frac{1/4}{1/2} = \frac{1}{2}, \quad \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(A \cup B)} = \frac{1/4}{3/4} = 1/3.$$

A generalization of Alice's reasoning is that if A , B , and C are events such that $B \subset C$ and $A \cap B = A \cap C$ (for example, if $A \subset B \subset C$), then the event A is at least as likely if we know that B has occurred than if we know that C has occurred. Alice's reasoning corresponds to the special case where $C = A \cup B$.

- (a) Each possible outcome has probability $1/36$. There are 6 possible outcomes that are doubles, so the probability of doubles is $6/36 = 1/6$.
 (b) The conditioning event (sum is 4 or less) consists of the 6 outcomes

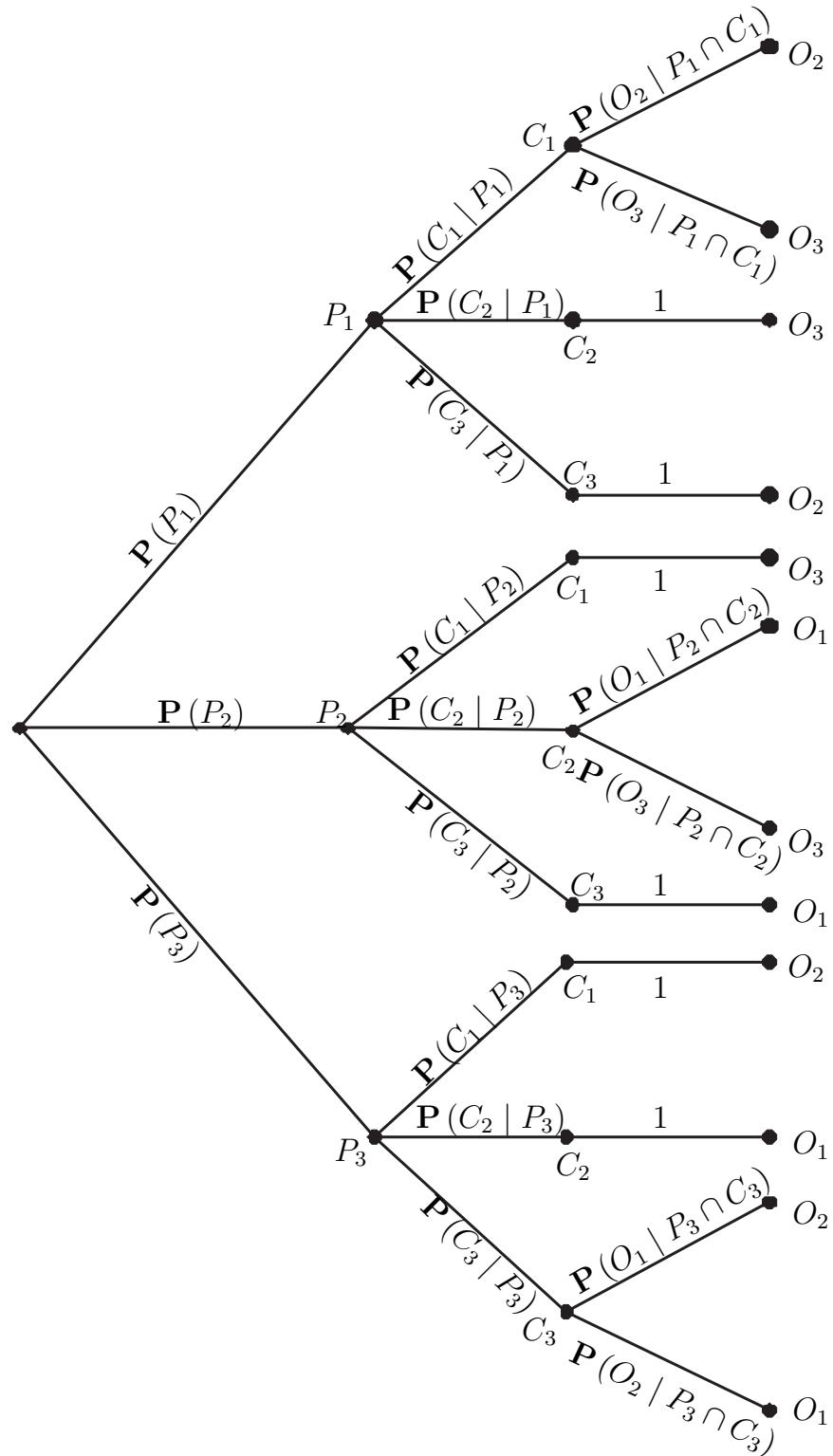
$$\{(1, 1), (1, 2), (1, 3), (2, 1), (2, 2), (3, 1)\},$$

2 of which are doubles, so the conditional probability of doubles is $2/6 = 1/3$.

- (c) There are 11 possible outcomes with at least one 6, namely, $(6, 6)$, $(6, i)$, and $(i, 6)$, for $i = 1, 2, \dots, 5$. Thus, the probability that at least one die is a 6 is $11/36$.
 (d) There are 30 possible outcomes where the dice land on different numbers. Out of these, there are 10 outcomes in which at least one of the rolls is a 6. Thus, the desired conditional probability is $10/30 = 1/3$.
- (a) See the textbook, Example 1.13, page 29.
 (b) See the textbook, Example 1.17, page 33.
- See the textbook, Example 1.12 (The Monty Hall Problem), page 27.

An alternative solution is given below:

Let P_i denote the event where the prize is behind door i , C_i denote the event where you initially choose door i , and O_i denote the event where your friend opens door i . The corresponding probability tree is:



MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
 (Fall 2010)

- (a) The probability of winning when not switching from your initial choice is the probability that the prize is behind the door you initially chose:

$$\begin{aligned}
 \mathbf{P}(\text{Win when not switching}) &= \mathbf{P}(P_1 \cap C_1) + \mathbf{P}(P_2 \cap C_2) + \mathbf{P}(P_3 \cap C_3) \\
 &= \mathbf{P}(P_1)\mathbf{P}(C_1|P_1) + \mathbf{P}(P_2)\mathbf{P}(C_2|P_2) + \mathbf{P}(P_3)\mathbf{P}(C_3|P_3) \\
 &= \mathbf{P}(P_1)\mathbf{P}(C_1) + \mathbf{P}(P_2)\mathbf{P}(C_2) + \mathbf{P}(P_3)\mathbf{P}(C_3) \\
 &= 1/3 \cdot (\mathbf{P}(C_1) + \mathbf{P}(C_2) + \mathbf{P}(C_3)) \\
 &= 1/3
 \end{aligned}$$

- (b) The probability of winning when switching from your initial choice is the probability that the prize is behind the remaining (unopened) door:

$$\begin{aligned}
 \mathbf{P}(\text{Win when switching}) &= \mathbf{P}(P_1 \cap C_2 \cap O_3) + \mathbf{P}(P_1 \cap C_3 \cap O_2) + \mathbf{P}(P_2 \cap C_1 \cap O_3) \\
 &\quad + \mathbf{P}(P_2 \cap C_3 \cap O_1) + \mathbf{P}(P_3 \cap C_1 \cap O_2) + \mathbf{P}(P_3 \cap C_2 \cap O_1) \\
 &= \mathbf{P}(P_1 \cap C_2) + \mathbf{P}(P_1 \cap C_3) + \mathbf{P}(P_2 \cap C_1) + \mathbf{P}(P_2 \cap C_3) \\
 &\quad + \mathbf{P}(P_3 \cap C_1) + \mathbf{P}(P_3 \cap C_2) \\
 &= \mathbf{P}(P_1)\mathbf{P}(C_2) + \mathbf{P}(P_1)\mathbf{P}(C_3) + \mathbf{P}(P_2)\mathbf{P}(C_1) + \mathbf{P}(P_2)\mathbf{P}(C_3) \\
 &\quad + \mathbf{P}(P_3)\mathbf{P}(C_1) + \mathbf{P}(P_3)\mathbf{P}(C_2) \\
 &= 2/3 \cdot (\mathbf{P}(C_1) + \mathbf{P}(C_2) + \mathbf{P}(C_3)) \\
 &= 2/3
 \end{aligned}$$

- (c) Given C_1 , that you first choose door 1, with the new strategy of switching only if door 3 is opened, you win if the prize behind door 1 and door 2 is opened or if the prize is behind door 2 and door 3 is opened.

$$\begin{aligned}
 \mathbf{P}(\text{Win with new strategy}|C_1) &= \mathbf{P}(P_1 \cap O_2|C_1) + \mathbf{P}(P_2 \cap O_3|C_1) \\
 &= \mathbf{P}(P_1|C_1)\mathbf{P}(O_2|P_1 \cap C_1) + \mathbf{P}(P_2|C_1)\mathbf{P}(O_3|P_2 \cap C_1) \\
 &= \mathbf{P}(P_1)\mathbf{P}(O_2|P_1 \cap C_1) + \mathbf{P}(P_2)\mathbf{P}(O_3|P_2 \cap C_1) \\
 &= 1/3 \cdot \mathbf{P}(O_2|P_1 \cap C_1) + 1/3 \cdot 1 \\
 &= 1/3 \cdot (\mathbf{P}(O_2|P_1 \cap C_1) + 1)
 \end{aligned}$$

Given that your initial choice is door 1, the probability of winning under this new strategy is dependent on how your friend decides which of doors 2 or 3 to open if the prize also lies behind door 1. If he always picks door 2, then $\mathbf{P}(O_2|P_1 \cap C_1) = 1$ and $\mathbf{P}(\text{Win with new strategy}|C_1) = 2/3$. If he picks between doors 2 and 3 with equal probability then $\mathbf{P}(O_2|P_1 \cap C_1) = 1/2$ and $\mathbf{P}(\text{Win with new strategy}|C_1) = 1/2$.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013 Transcript – Recitation: A Chess Tournament Problem

Hi. Welcome back. Today, we're going to do a fun problem called the chess tournament problem. Now, it's a very long problem, so I just want to jump straight in.

Essentially, the problem statement describes a very special chess tournament, which involves players named Al, Bo, and Chi. Now Al is the current reigning champion, and Bo and Chi are this year's contenders, and, of course, they're vying with each other to beat out Al and become the new champion. And so essentially, the tournament is divided into two rounds-- a first round, during which Bo and Chi play against each other, and then a second round, during which the surviving contender from the first round plays against Al. And the problem statement also gives you a bunch of information like what's the probability that Bo beats Chi in a particular game, et cetera.

So without further ado, let's get started on part a. In part a, the first thing we're asked to compute is the probability that a second round is required. Now, to save myself some writing, I used the notation R2 to denote that event. So we are interested in probability of R2.

Now, I claim that this problem is very sequential in nature so I would like to draw a tree to describe what's happening. So in the first part of the tournament, when Bo and Chi play their first game, exactly one of two things can happen-- either Bo can win or Chi can win. And we're told by the problem statement that Bo wins with the probability of 0.6 and, therefore, Chi must win with the probability of 0.4, right? Because these two possibilities must sum to 1, because either this must happen or this happen.

Now, let's imagine that the first game has been played and that Bo won. Well, during the second game, there's still two options for the outcome-- Bo could win the second game or Chi could win the second game. And because the problem statement says that in every scenario Bo always wins against Chi with the probability of 0.6, we can go ahead and put a 0.6 along this branch as well. Similarly, 0.4 here. And similar logic, you've got a tree that looks like this.

And for those of you who haven't seen trees before, it's just a structure that looks something like this. And it helps us do better accounting. It helps us keep straight in our head what are the various outcomes, so that we don't get confused. And so very quickly here, you can see that there's four possible outcomes.

So each node in this tree corresponds to an outcome. And the leaves are those nodes at the furthest stage. And it's convention to draw the probability of a particular-- so for instance, the probability that Bo wins the first game-- it's just convention to draw that probability over the corresponding branch. And the reason why such diagrams are so useful is because to compute the probability of a particular outcome, if you've designed your tree correctly, all you have to do is multiply the probabilities along the branches that get into that outcome.

So let's see that in action. When is a second round required? Well, a second round is required here, right? Because in this case, Bo would be the surviving challenger and he'd play the next round against Al.

It's also required here. But of course, it's not required here or here, because no second round is played. And so these two outcomes comprise the event R2.

And now, to get the probability of this outcome, you multiply along the branches. So 0.6 times 0.6 give you 0.36. And 0.4 times 0.4 gives you 0.16. And we're almost done.

We know that these two events are disjoint, because if Bo won the first two games, then, certainly, Chi could've won the first two games. And so you can just sum the probabilities to get the probability of the reunion. So the probability of R2 is equal to the probability that Bo won the first two games or Chi won the first two games. And that's equal to 0.36 plus 0.16, which is equal to a 0.52.

OK, now the second part of part a asks for the probability that Bo wins the first round. And this is first round. This is a very straightforward one. So Bo wins the first round, that corresponds only to this particular outcome. And we already know the probability associated with that outcome is equal to the 0.36. So we're done with that one.

And now the last part is sort of an interesting one. It asks for the probability that Al retains his championship this year. So I'm going to just call that A for short. A is the event that Al retains his championship this year. And for that we're going to need a larger tree, because Al has a lot of activity in the second round, and so far our tree only describes what happens in the first round.

Now, to save time, I've actually drawn the rest of the tree over there up in the corner. So let's get rid of this one and let's look at the full tree. So let's see when does Al retain his championship? Well, Al certainly retains his championship here, right? Because no second round is required. Similarly, here.

Al retains his championship here, because the second round was required, but Al beat Bo. And similarly, here Bo didn't win both games in the second round against Al, so Al wins. Here, Bo is the new champion. So we don't want to include about one. And sort of by symmetry, we also get this one and this one.

So by my argument before, we know that the outcomes that comprise our event of interest are this one, this one, this one, this one, this one, and this one. So we could multiply the probabilities along each branch and sum them, because they're disjoint, to get the total probability. But we're not going to do that because that's a lot of algebra. Instead, we're going to look at the complement of the event.

So we're going to notice, there's only two branches on which Al does not retain his current championship. So P of A is, of course, equal to 1 minus P of A. And we're going to get P of A by inspection. I'm sorry, so P of A compliment. I'm just testing you, guys.

So P of A compliment corresponds to here and to here, because those are the outcomes where $A1$ didn't win. And so again, you multiply along the branches to get the probabilities. So you get 0.6 squared times 0.5 squared plus 0.4 squared times 0.3 squared. And if you do all the algebra, you should get around 0.8956 .

So we're cruising through this problem. Let's go to part b. Part b is a little bit less straightforward than part a, because it starts asking you for conditional probabilities, as opposed to a priori probabilities. So in the first part-- and again, I'm going to continue my notation with $R2$ -- we want the probability that $B0$ is the surviving challenger-- so I'm just going to use B to denote that-- given $R2$.

Now, by definition, you should remember from lecture that this is equal to probability of B and $R2$ divided by the probability of $R2$. And of course, we've already computed this value right up here with part a. We know it's 2.5 .

So we don't have to do any more work there. We only have to look at the numerator. So we need to go and figure out what nodes in that tree correspond to the event B intersect $R2$.

So let's use a new color. Let's see, $B0$ is the surviving challenger here only, right? And $R2$ is automatically satisfied, right? Because a second round is required there and there, not on those two. But here Chi is the surviving challenger, not $B0$, so we're really only interested in that node. And you multiply along the branches to get the probabilities. So we have 0.36 over 0.52 , which is approximately equal to 0.6923 .

OK, now, the next part wants the conditional probability that AL retains his championship, conditioned, again, on $R2$. So we already have A being the event $A1$ retains his championship. So we want the probability of A , given $R2$.

And let's just apply the direct definition of conditional probability again. You get P of A and $R2$ divided by a probability of $R2$. Of course, we have the probability of $R2$ already, so we just need to find the node in the tree that corresponds to A and $R2$.

So where is $R2$? $R2$ is going to correspond to every node to the right that is not one of these two. So a second round is required here, here, here, here, here, and here.

Now, where does $A1$ retain his championship? So $A1$ retains his championship here. He retains his championship here. He retains his championship here and here, but no second round is required, so these guys don't belong in the intersection. But this does, and this does. So we can again multiply the probabilities along the branches and then some them.

So let's see, we get-- this marker's not working very well, so I'm going to switch back to the pink-- so you get 0.6 squared times 0.5 . That gets rid of this one. And then we want 0.6 squared times 0.5 squared. That gets rid of that one.

And then plus-- let's see-- 0.4 squared times 0.7 , which takes care of this one. And then lastly, 0.4 squared times 0.3 times 0.7 . And that is a long expression. But it happens to be about 0.7992 .

OK, so we are done with part b and we can move along to part c. And I am, since we're running out of room, I'm actually just going to erase this. And hopefully you guys have had a chance to copy it down by now. If not, you can always pause the video and go back.

So let's see, part c asks us given that the second round is required and that it comprised of one game only. So let's denote I. So let's I be the event that the second round was one game only. So essentially, in math conditioned on R2 and I, what is the probability that it was Bo who won the first round?

So let's let B be the event that Bo won the first round. OK, so again translating the English to math, we just want the probability of B given R2 and I. Now, I am once again going to use the definition of conditional probability.

You might be concerned that we haven't defined explicitly yet the definition of conditional probability, when what lies behind the conditioning bar is not a single event, but it's rather an intersection of an event. And so my claim to you is that it doesn't matter and that the same exact definition applies. But we'll go through it slowly.

So R2 is an event, I is an event, and we know that the intersection of two events is itself an event. So I'm going to make up a new letter, and I'm going to call this event W. So just using the new notation, this is equal to probability of B, given W.

Now, this is the normal definition that we know. We know that this is probability of B intersect W over probability of W. And then we just resubstitute what the definition of W was. And so if you do that over here, you get probability of B and R2 and I divided by probability of R2 and I.

So hopefully, jumping from here ahead to here, you see that the definitions act exactly the same way. But these are two very short intermediate steps that should help you convince yourself that the same definition still works. So let's start with the denominator, because the denominator looks a little bit easier. Where is R2 and I in our tree?

Well, let's see. Here, a second round was required, but it comprised two games. Same with this one. Here, a second round was required and it was comprised only of one game. So this is good. This is one of the outcomes that we're looking for.

Here, no second round was required. So this doesn't qualify. Same with this one. Here, a second round was required, and there was only one game, so that's good. And then these don't qualify for the same reasons as we set up there.

So we just have to multiply the probabilities along those branches. And we see that it's 0.4 squared times 0.7 plus 0.6 squared times 0.5.

OK, we're almost done. We just need to look at the intersection of R2 and I. So R2 and I are the ones we've already circled. But now, we want to add one more constraint, which is that Bo had to have won the first round.

And so we see here that Chi won the first round, if we're looking at this outcome. And so he's no good. Let's use a different color. Let's see, maybe this one. But here Bo did win the first round.

So we're going to get 0.6 squared times 0.5 . And I got that, of course, just by multiplying the probabilities along the right branches. And this, if you're curious, comes out to be about 0.6164 .

OK, so I know that was a lengthy problem, but you should feel really comfortable now doing sort of basic probability manipulations. One thing that this problem emphasized a lot was your ability to compute conditional probabilities. So you saw me apply the definition of conditional probability twice in part b. And then you saw me apply the definition again in part c in a sort of slightly modified way. So that's one thing that you should have gotten out of this problem.

And then another thing is that hopefully, you noticed that by using a tree diagram, we made the problem much easier. We almost didn't even have to think about computing probabilities anymore. We reduced the problem to just saying, OK, what are the outcomes that comprise our event of interest? And then once you select those, to compute their probability you multiply the probabilities along the branches.

You have the right to just add those together, because if you draw your tree correctly, all of these guys should be disjoint from one another. So you have to be careful, of course, to set up your tree appropriately. But once you do set up your tree appropriately, your life is much simpler. So that's it for today. And we'll see you next time.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013

Transcript – Recitation: A Random Walker

In this problem, we'll be working with a object called random walk, where we have a person on the line-- or a tight rope, according to the problem. Let's start from the origin, and each time step, it would randomly either go forward or backward with certain probability. In our case, with probability P , the person would go forward, and $1 - P$ going backwards.

Now, the walk is random in the following sense-- that the choice going forward or backward in each step is random, and it's completely independent from all past history. So let's look at the problem. It has three parts. In the first part, we'd like to know what's the probability that after two steps the person returns to the starting point, which in this case is 0?

Now, throughout this problem, I'm going to be using the following notation. F indicates the action of going forward and B indicates the action of going backwards. A sequence says F and B implies the sample that the person first goes forward, and then backwards.

If I add another F , it will mean, forward, backward, forward again. OK? So in order for the person to go somewhere after two steps and return to the origin, the following must happen. Either the person went forward followed by backward, or backward followed by forward.

And indeed, this event-- namely, the union of these two possibilities-- defines the event of interest in our case. And we'd like to know what's the probability of A , which we'll break down into the probability of forward, backward, backward, forward.

Now, forward, backward and backward, forward-- they are two completely different outcomes. And we know that because they're disjoint, this would just be the sum of the two probabilities-- plus probability of backward/forward. Here's where the independence will come in.

When we try to compute the probability of going forward and backward, because the action-- each step is completely independent from the past, we know this is the same as saying, in the first step, we have probability P of going forward, in the next step, probability $1 - P$ of going backwards.

We can do so-- namely, writing the probability of forward, backward as a product of going forward times the probability of going backwards, because these actions are independent. And similarly, for the second one, we have going backwards first, times going forward the second time.

Adding these two up, we have $2 \times P \times (1 - P)$. And that will be the answer to the first part of the problem. In the second part of the problem, we're interested in the probability that

after three steps, the person ends up in position 1, or one step forward compared to where he started.

Now, the only possibilities here are that among the three steps, exactly two steps are forward, and one step is backwards, because otherwise there's no way the person will end up in position 1. To do so, there, again, are three possibilities in which we go forward, forward, backward, or forward, backward, forward, or backward, forward.

And that exhausts all the possibilities that the person can end up in position 1 after three steps. And we'll define the collection of all these outcomes as event C. The probability of event C-- same as before-- is simply the sum of the probability of each individual outcome.

Now, based on the independence assumption that we used before, each outcome here has the same probability, which is equal to P squared times $1 - P$. The P squared comes from the fact that two forward steps are taken, and $1 - P$, the probability of that one backwards step.

And since there are three of them, we multiply 3 in front, and that will give us the probability. In the last part of the problem, we're asked to compute that, conditional on event C already took place, what is the probability that the first step he took was a forward step?

Without going into the details, let's take a look at the C, in which we have three elements, and only the first two elements correspond to a forward step in the first step. So we can define event D as simply the first two outcomes-- forward, forward, backward, and forward, backward, forward.

Now, the probability we're interested in is simply probability of D conditional on C. We'd write it out using the law of conditional probability-- D intersection C conditional on C. Now, because D is a subset of C, we have probability of D divided by the probability of C.

Again, because all samples here have the same probability, all we need to do is to count the number of samples here, which is 2, and divide by the number of samples here, which is 3. So we end up with 2 over 3. And that concludes the problem. See you next time.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013 Transcript – Recitation: Communication over a Noisy Channel

Hi. In this problem, we'll be talking about communication across a noisy channel. But before we dive into the problem itself, I wanted to first motivate the context a little bit and talk more about what exactly a communication channel is and what "noise" means. So in our everyday life, we deal with a lot of communication channels, for example, the internet, where we download data and we watch videos online, or even just talking to a friend. And the air could be your communication channel for our voice.

But as you probably have experienced, sometimes these channels have noise, which just means that what the sender was trying to send isn't necessarily exactly what the receiver receives. And so in probability, we try to model these communication channels and noise and try to understand the probability behind it. And so now, let's go into the problem itself.

In this problem, we're dealing with a pretty simple communication channel. It's just a binary channel, which means that what we're sending is just one bit at a time. And here, a bit just means either 0 or 1-- so essentially, the simplest case of information that you could send. But sometimes when you send a 0, the receiver actually receives a 1 instead, or vice versa.

And that is where noise comes in. So here in this problem, we actually have a probabilistic model of this channel and the noise that hits the channel. What we're trying to send is either 0 or a 1. And what we know is that on the receiving end, a 0 can either be received when a 0 is sent, or a 1 can be received when a 0 is sent.

And when a 1 is sent, we could also have noise that corrupts it. And you get a 0 instead. Or you can have a 1 being successfully transmitted.

And the problem actually tells us what the probabilities here are. So we know that if a 0 is sent, then with probability $1 - \epsilon_0$, a 0 is received. And with the remaining probability, it actually gets corrupted and turned into a 1. And similarly, if a 1 is sent, then with probability $1 - \epsilon_1$, the 1 is correctly transmitted. And with the remaining probability ϵ_1 , it's turned into a 0 instead.

And the last bit of information is that we know that with the probability p , any random bit is actually 0 that is being sent. And with probability $1 - p$, we're actually trying to send a 1. So that is the basic setup for the problem.

And the first part that the problem asks us to find, what is the probability of a successful transmission when you have just any arbitrary bit that's being sent. So what we can do here is, use this tree that we've already drawn and identify what are the cases, the outcomes where a bit is actually successfully transmitted. So if a 0 is sent and a 0 is received, then that corresponds to a successful transmission.

Similarly, if a 1 is sent and a 1 is received, that also corresponds to a successful transmission. And then we can calculate what these probabilities are, because we just calculate the probabilities along the branches. And so here implicitly, what we're doing is invoking the multiplication rule.

So we can calculate the probabilities of these two individual outcomes and their disjoint outcomes. So we can actually just sum the two probabilities to find the answer. So the probability here is p times 1 minus ϵ_0 -- that's the probability of a 0 being successfully transmitted -- plus 1 minus p times 1 minus ϵ_1 , which is the probability that a 1 is successfully transmitted.

And so what we've done here is actually just looked at this kind of diagram, this tree to find the answer. What we also could have done is been a little bit more methodical maybe and actually apply the law of total probability, which is really what we're trying to do here. So you can see that this actually corresponds to -- the p corresponds to the probability of 0 being sent. And 1 minus ϵ_0 is the probability of success, given that a 0 is sent.

And this second term is analogous. It's the probability that a 1 was sent times the probability that you have a success, given that a 1 was sent. And this is just an example of applying the law of total probability, where we partitioned into the two cases of a 0 being sent and a 1 being sent and calculated the probabilities for each of those two cases and added those up.

So that's kind of a review of the multiplication rule and law of total probability. So now, let's move on to part B. Part B is asking, what is the probability that a particular sequence of bits, not just a single one, but a sequence of four bits in a row is successfully transmitted? And the sequence that we're looking for is, 1, 0, 1, 1. So this is how I'll denote this event.

1, 0, 1, 1 gets successfully transmitted into 1, 0, 1, 1. Now, instead of dealing with single bits in isolation, we have a sequence of four bits. But we can really just break this out into the four individual bits and look at those one by one.

So in order to transmit successfully 1, 0, 1, 1, that whole sequence, we first need to transmit a 1 successfully, then a 0 successfully, then another 1 successfully, and then finally, the last 1 successfully. So really, this is the same as the intersection of four different smaller events, a 1 being successfully transmitted and a 0 being successfully transmitted and two more 1's being successfully transmitted.

So why are we able to do this, first of all? We are using an important assumption that we make in the problem that each transmission of an individual bit has the same probabilistic structure so that no matter which bit you're talking about, they all have the same [? error ?] probability, the same probabilities of being either successfully transmitted or having noise corrupt it.

So because of that, it doesn't really matter which particular 1 or 0 we're talking about. And now, we'll make one more step, and we'll invoke independence, which is the third topic here. And the other important assumption here we're making is that every single bit is independent from any

other bit. So the fact that this one was successfully transmitted has no impact on the probability of the 0 being successfully transmitted or not.

And so because of that, we can now break this down into a product of four probabilities. So this becomes the probability of 1 transmitted into a 1 times the probability of 0 transmitted into a 0, 1 to a 1, and 1 to 1. And that simplifies things, because we know what each one of these are. The probability of 1 being successful transmitted into a 1, we know that's just 1 minus epsilon 1.

And similarly, probability of 0 being transmitted into a 0 is 1 minus epsilon naught. So our final answer then is just-- well, we have three of these and one of these. So the answer is going to be 1 minus epsilon naught times 1 minus epsilon 1 to the third power.

Now, let's move on go on to part C, which adds another wrinkle to the problem. So now, maybe we're not satisfied with the success rate of our current channel. And we want to improve it somehow. And one way of doing this is to add some redundancy. So instead of just sending a single 0 and hoping that it gets successfully transmitted, instead what we can do is, send three 0's in a row to represent a single 0 and hope that because we've added some redundancy, we can somehow improve our error rates.

So in particular what we're going to do is, for a 0, when we want to send a 0, which I'll put in quotes here, what we're actually going to send is a sequence of three 0s. And what's going to happen is, this sequence of three 0s, each one of these bits is going to go through the same channel. So the 0, 0, 0 can stay and get transmitted successfully as a 0, 0, 0.

Or maybe the last 0 gets turned into a 1, or the second 0 gets turned into a 1, or we can have any one of these eight possible outcomes on the receiving end. And then similarly, for a 1, when we want to send a 1, what we'll actually send is a sequence of three 1's, three bits. And just like above, this 1, 1, 1, due to the noise in the channel, it can get turned into any one of these eight sequences on the receiving end.

So what we're going to do now is, instead of sending just a single 0, we'll send three 0s, and instead of sending a 1, we'll send three 1s. But now, the question is, this is what you'll get on the receiving end. How do you interpret-- 0, 0, 0, maybe intuitively you'll say that's obviously a 0.

But what if you get something like 0, 1, 0 or 1, 0, 1, when there's both 0s and 1s in the received message? What are you going to do? So one obvious thing to do is to take a majority rule. So because there's three of them, if there's two or more 0s, we'll say that what was meant to be sent was actually a 0. And if there's two or more 1s, then we'll interpret that as a 1 being sent.

So in this case, let's look at the case of 0. The majority rule here would say that, well, if 0, 0, 0 was sent, then the majority is 0s. And similarly, in these two cases, 0, 0, 1 or 0, 1, 0, the majority is also 0s. And then finally, in this last case, 1, 0, 0, you get a majority of 0s.

So in these four received messages, we'll interpret that as a 0 have been set. So part C is asking, given this majority rule and this redundancy, what is the probability that a 0 is correctly transmitted? Well, to answer that, we've already identified these are the four outcomes, where a 0

would be correctly transmitted. So to find the answer to this question, all we have to do is find the probability that a sequence of 0, 0, 0 gets turned into one of these four sequences.

So let's do that. What is the probability that a 0, 0, 0 gets turned into a 0, 0, 0? Well, that means that all three of these 0s had no errors. So we would have the answer being 1 minus epsilon 0 cubed, because all three of these bits had to have been successfully transmitted.

Now, let's consider the other ones. For example, what's the probability that a 0, 0, 0 gets turned into a 0, 0, 1? Well, in this case, we need two successful transmissions of 0s, plus one transmission of 0 that had an error.

So that is going to be 1 minus epsilon naught squared for the two successful transmissions of 0, times epsilon naught for the single one that was wrong. And if you think about it, that was only for this case-- 0, 0, 1. But the case where 0, 1, 0 and 1, 0, 0 are the same, because for all three of these, you have two successful transmissions of 0, plus one that was corrupted with noise.

And so it turns out that all three of those probabilities are going to be the same. So this is our final answer for this part. Now, let's move on to part D. Part D is asking now a type of inference problem. And we'll talk more about inference later on in this course.

The purpose of this problem-- what it's asking is, suppose you received a 1, 0, 1. That's the sequence of three messages, three bits that you received. Given that you received a 1, 0, 1, what's the probability that 0 was actually the thing that was being sent.

So if you look at this, you'll look at it and say, this looks like something where we can apply Bayes' rule. So that's the fourth thing that we're covering in this problem. And if you apply Bayes' rule, what you'll get is, this is equal to the probability of 0 times the probability of 1, 0, 1 being received, given that 0 was what was sent, divided by the probability that 1, 0, 1 is received.

So we have this basic structure. And we also know that we can use the law of total probability again on this denominator. So we know that the probability that 1, 0, 1 is received is equal to the probability of 0 being sent times probability of 1, 0, 1 being received, given that 0 was sent, plus the probability that 1 was sent times the probability that 1, 0, 1 is received, given that 1 is sent.

And as you'll notice in applications of Bayes' rule, usually what you'll have is a numerator is then repeated as one of the terms in the denominator, because it's just an application of total probability. So if you put these pieces together, really, what we need is just these four terms. Once we have those four terms, we can just plug them into this equation, and we'll have our answer.

So let's figure out what those four terms are. The probability of 0 being sent-- well, we said that earlier. Probability of 0 being sent is just p . And the probability of 1 being sent is $1 - p$. That's just from the model that we're given in the problem. Now, let's figure out this part.

What is the probability of a 1, 0, 1 being received, given that 0 was sent? So if 0 was sent, then we know that what really was sent was 0, 0, 0, that sequence of three bits. And now, what's the probability that 0, 0, 0 got turned into 1, 0, 1?

Well, in this case, what we have is one successful transmission of a 0, plus two failed transmission of a 0. So that one successful transmission of a 0, that probability is 1 minus epsilon naught. And now, we have two failed transmissions of a 0.

So we have to multiply that by epsilon naught squared. And now, for the last piece, what's the probability of receiving the 1, 0, 1, given that 1 was actually sent? Well, in that case, if a 1 was sent, what was really sent was a sequence of three 1s. And now, we want the probability that a 1, 1, 1 got turned into a 1, 0, 1.

In this case, we have two successful transmissions of the 1 with one failed transmission. So the two successful transmissions will have 1 minus epsilon 1 squared. And then the one failed transmission will give us an extra term of epsilon 1. So just for completeness, let's actually write out what this final answer would be.

So probability of 0 is p. Probability of 1, 0, 1, given 0 is, we calculated that as 1 minus epsilon naught times epsilon naught squared. The same term appears again in the denominator.

Plus the other term is, probability of 1 times the probability of 1, 0, 1, given 1. So that is 1 minus epsilon squared times epsilon 1. So that is our final answer. And it's really just a application of Bayes' rule.

So this was a nice problem, because it represents a real world phenomenon that happens. And we can see that you can apply a pretty simple probabilistic model to it and still be able to answer some interesting questions. And there are other extensions that you can ask also. For example, we've talked about adding redundancy by tripling the number of bits, but tripling the number of bits also reduces the throughputs, because instead of sending one, you have to send three bits just to send one.

So if there's a cost of that, at what point does the benefit of having lower error outweigh the cost of having to send more things? And so that's a question that you can answer with some more tools in probability. So we hope you enjoyed this problem. And we'll see you again next time.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013

Transcript – Lecture 3

The following content is provided under a Creative Commons license. Your support will help MIT OpenCourseWare continue to offer high quality educational resources for free. To make a donation or view additional materials from hundreds of MIT courses, visit MIT OpenCourseWare at ocw.mit.edu.

PROFESSOR: Let us start. So as always, we're to have a quick review of what we discussed last time. And then today we're going to introduce just one new concept, the notion of independence of two events. And we will play with that concept.

So what did we talk about last time? The idea is that we have an experiment, and the experiment has a sample space Ω . And then somebody comes and tells us you know the outcome of the experiments happens to lie inside this particular event B . Given this information, it kind of changes what we know about the situation. It tells us that the outcome is going to be somewhere inside here. So this is essentially our new sample space.

And now we need to we reassign probabilities to the various possible outcomes, because, for example, these outcomes, even if they had positive probability beforehand, now that we're told that B occurred, those outcomes out there are going to have zero probability. So we need to revise our probabilities. The new probabilities are called conditional probabilities, and they're defined this way.

The conditional probability that A occurs given that we're told that B occurred is calculated by this formula, which tells us the following-- out of the total probability that was initially assigned to the event B , what fraction of that probability is assigned to outcomes that also make A to happen? So out of the total probability assigned to B , we see what fraction of that total probability is assigned to those elements here that will also make A happen.

Conditional probabilities are left undefined if the denominator here is zero. An easy consequence of the definition is if we bring that term to the other side, then we can find the probability of two things happening by taking the probability that the first thing happens, and then, given that the first thing happened, the conditional probability that the second one happens.

Then we saw last time that we can divide and conquer in calculating probabilities of mildly complicated events by breaking it down into different scenarios. So event B can happen in two ways. It can happen either together with A , which is this probability, or it can happen together with A complement, which is this probability. So basically what we're saying that the total probability of B is the probability of this, which is A intersection B , plus the probability of that, which is A complement intersection B .

So these two facts here, multiplication rule and the total probability theorem, are basic tools that one uses to break down probability calculations into a simpler parts. So we find probabilities of two things happening by looking at each one at a time. And this is what we do to break up a situation with two different possible scenarios.

Then we also have the Bayes rule, which does the following. Given a model that has conditional probabilities of this kind, the Bayes rule allows us to calculate conditional probabilities in which the events appear in different order. You can think of these probabilities as describing a causal model of a certain situation, whereas these are the probabilities that you get after you do some inference based on the information that you have available.

Now the Bayes rule, we derived it, and it's a trivial half-line calculation. But it underlies lots and lots of useful things in the real world. We had the radar example last time. You can think of more complicated situations in which there's a bunch or lots of different hypotheses about the environment. Given any particular setting in the environment, you have a measuring device that can produce many different outcomes. And you observe the final outcome out of your measuring device, and you're trying to guess which particular branch occurred. That is, you're trying to guess the state of the world based on a particular measurement.

That's what inference is all about. So real world problems only differ from the simple example that we saw last time in that this kind of tree is a little more complicated. You might have infinitely many possible outcomes here and so on. So setting up the model may be more elaborate, but the basic calculation that's done based on the Bayes rule is essentially the same as the one that we saw.

Now something that we discuss is that sometimes we use conditional probabilities to describe models, and let's do this by looking at a model where we toss a coin three times. And how do we use conditional probabilities to describe the situation? So we have one experiment. But that one experiment consists of three consecutive coin tosses. So the possible outcomes, our sample space, consists of strings of length 3 that tell us whether we had heads, tails, and in what sequence. So three heads in a row is one particular outcome.

So what is the meaning of those labels in front of the branches? So this P here, of course, stands for the probability that the first toss resulted in heads. And let me use this notation to denote that the first was heads. I put an H in toss one.

How about the meaning of this probability here? Well the meaning of this probability is a conditional one. It's the conditional probability that the second toss resulted in heads, given that the first one resulted in heads. And similarly this label here corresponds to the probability that the third toss resulted in heads, given that the first one and the second one resulted in heads. So in this particular model that I wrote down here, those probabilities, P , of obtaining heads remain the same no matter what happened in the previous toss.

For example, even if the first toss was tails, we still have the same probability, P , that the second one is heads, given that the first one was tails. So we're assuming that no matter what happened in the first toss, the second toss will still have a conditional probability equal to P . So that conditional probability does not depend on what happened in the first toss. And we will see that this is a very special situation, and that's really the concept of independence that we are going to introduce shortly.

But before we get to independence, let's practice once more the three skills that we covered last time in this example. So first skill was multiplication rule. How do you find the probability of several things happening? That is the probability that we have tails followed by heads followed by tails. So here we're talking about this particular outcome here, tails followed by heads followed by tails. And the way we calculate such a probability is by multiplying conditional probabilities along the path that takes us to this outcome. And so these conditional probabilities are recorded here. So it's going to be $(1 - P)$ times P times $(1 - P)$. So this is the multiplication rule.

Second question is how do we find the probability of a mildly complicated event? So the event of interest here that I wrote down is the probability that in the three tosses, we had a total of one head. Exactly one head. This is an event that can happen in multiple ways. It happens here. It happens here. And it also happens here. So we want to find the total probability of the event consisting of these three outcomes. What do we do? We just add the probabilities of each individual outcome. How do we find the probability of an individual outcome? Well, that's what we just did.

Now notice that this outcome has probability P times $(1 - P)$ squared. That one should not be there. So where is it? Ah. It's this one.

OK, so the probability of this outcome is $(1 - P)$ times P times $(1 - P)$, the same probability. And finally, this one is again $(1 - P)$ squared times P .

So this event of one head can happen in three ways. And each one of those three ways has the same probability of occurring. And this is the answer.

And finally, the last thing that we learned how to do is to use the Bayes rule to calculate and make an inference. So somebody tells you that there was exactly one head in your three tosses. What is the probability that the first toss resulted in heads? OK, I guess you can guess the answer here if I tell you that there were three tosses. One of them was heads. Where was that head in the first, the second, or the third?

Well, by symmetry, they should all be equally likely. So there should be probably just $1/3$ that that head occurred in the first toss. Let's check our intuition using the definitions. So the definition of conditional probability tells us the conditional probability is the probability of both things happening. First toss is heads, and we have exactly one head divided by the probability of one head.

What is the probability that the first toss is heads, and we have exactly one head? This is the same as the event heads, tails, tails. If I tell you that the first is heads, and there's only one head, it means that the others are tails. So this is the probability of heads, tails, tails divided by the probability of one head. And we know all of these quantities probability of heads, tails, tails is P times $(1 - P)$ squared. Probability of one head is 3 times P times $(1 - P)$ squared. So the final answer is $1/3$, which is what you should have guessed on intuitive grounds.

Very good. So we got our practice on the material that we did cover last time. Again, think. There's basically three basic skills that we are practicing and exercising here. In the problems, quizzes, and in the real life, you may have to apply those three skills in somewhat more complicated settings, but in the end that's what it boils down to usually.

Now let's focus on this special feature of this particular model that I discussed a little earlier. Think of the event heads in the second toss. Initially, the probability of heads in the second toss, you know, that it's P , the probability of success of your coin. If I tell you that the first toss resulted in heads, what's the probability that the second toss is heads? It's again P . If I tell you that the first toss was tails, what's the probability that the second toss is heads? It's again P . So whether I tell you the result of the first toss, or I don't tell you, it doesn't make any difference to you. You would always say the probability of heads in the second toss is going to P , no matter what happened in the first toss.

This is a special situation to which we're going to give a name, and we're going to call that property independence. Basically independence between two things stands for the fact that the first thing, whether it occurred or not, doesn't give you any information, does not cause you to change your beliefs about the second event. This is the intuition. Let's try to translate this into mathematics.

We have two events, and we're going to say that they're independent if your initial beliefs about B are not going to change if I tell you that A occurred. So you believe something how likely B is. Then somebody comes and tells you, you know, A has happened. Are you going to change your beliefs? No, I'm not going to change them. Whenever you are in such a situation, then you say that the two events are independent.

Intuitively, the fact that A occurred does not convey any information to you about the likelihood of event B . The information that A provides is not so useful, is not relevant. A has to do with something else. It's not useful for your guessing whether B is going to occur or not.

So we can take this as a first attempt into a definition of independence. Now remember that we have this property, the probability of two things happening is the probability of the first times the conditional probability of the second. If we have independence, this conditional probability is the same as the unconditional probability.

So if we have independence according to that definition, we get this property that you can find the probability of two things happening by just multiplying their individual probabilities. Probability of heads in the first toss is $1/2$. Probability of heads in the second toss is $1/2$. Probability of heads heads is $1/4$. That's what happens if your two tosses are independent of each other.

So this property here is a consequence of this definition, but it's actually nicer, better, simpler, cleaner, more beautiful to take this as our definition instead of that one. Are the two definitions equivalent? Well, they're almost the same, except for one thing. Conditional probabilities are only defined if you condition on an event that has positive probability.

So this definition would be limited to cases where event A has positive probability, whereas this definition is something that you can write down always. We will say that two events are independent if and only if their probability of happening simultaneously is equal to the product of their two individual probabilities.

And in particular, we can have events of zero probability. There's nothing wrong with that. If A has 0 probability, then A intersection B will also have zero probability, because it's an even smaller event. And so we're going to get zero is equal to zero. A corollary of what I just said, if an event A has zero probability, it's actually independent of any other event in our model, because we're going to get zero is equal to zero. And the definition is going to be satisfied.

This is a little bit harder to reconcile with the intuition we have about independence, but then again, it's part of the mathematical definition. So what I want you to retain is this notion that the independence is something that you can check formally using this definition, but also you can check intuitively by if, in some cases, you can reason that whatever happens and determines whether A is going to occur or not, has nothing absolutely to do with whatever happens and determines whether B is going to occur or not.

So if I'm doing a science experiment in this room, and it gets hit by some noise that's causes randomness. And then five years later, somebody somewhere else does the same science experiment somewhere else, it gets hit by other noise, you would usually say that these experiments are independent. So what events happen in one experiment are not going to change your beliefs about what might be happening in the other, because the sources of noise in these two experiments are completely unrelated. They have nothing to do with each other.

So if I flip a coin here today, and I flip a coin in my office tomorrow, one shouldn't affect the other. So the events that I get from these should be independent. So that's usually how independence arises. By having distinct physical phenomena that do not interact.

Sometimes you also get independence even though there is a physical interaction, but you just happen to have a numerical accident. A and B might be physically related very tightly, but a numerical accident happens and you get equality here, that's another case where we do get independence.

Now suppose that we have two events that are laid out like this. Are these two events independent or not? The picture kind of tells you that one is separate from the other. But separate has nothing to do with independent. In fact, these two events are as dependent as Siamese twins. Why is that?

If I tell you that A occurred, then you are certain that B did not occur. So information about the occurrence of A definitely affects your beliefs about the possible occurrence or non-occurrence of B. When the picture is like that, knowing that A occurred will change drastically my beliefs about B, because now I suddenly become certain that B did not occur.

So a picture like this is a case actually of extreme dependence. So don't confuse independence with disjointness. They're very different types of properties.

AUDIENCE: Question.

PROFESSOR: Yes?

AUDIENCE: So I understand the explanation, but the probability of A intersect B [INAUDIBLE] to zero, because they're disjoint.

PROFESSOR: Yes.

AUDIENCE: But then the product of probability A and probability B, one of them is going to be 1. [INAUDIBLE]

PROFESSOR: No, suppose that the probabilities are $1/3$, $1/4$, and the rest is out there. You check the definition of independence. Probability of A intersection B is zero. Probability of A times the probability of B is $1/12$. The two are not equal. Therefore we do not have independence.

AUDIENCE: Right. So what's wrong with the intuition of the probability of A being 1, and the other one being 0? [INAUDIBLE].

PROFESSOR: No. The probability of A given B is equal to 0. Probability of A is equal to $1/3$. So again, these two are different. So we had some initial beliefs about A, but as soon as we are told that B occurred, our beliefs about A changed. And so since our beliefs changed, that means that B conveys information about A.

AUDIENCE: So can you not draw independent [INAUDIBLE] on a Venn diagram?

PROFESSOR: I can't hear you.

AUDIENCE: Can you draw independence on a Venn diagram?

PROFESSOR: No, the Venn diagram is never enough to decide independence. So the typical picture in which you're going to have independence would be one event this way, and another event this way. You need to take the probability of this times the probability of that, and check that, numerically, it's equal to the probability of this intersection. So it's more than a Venn diagram. Numbers need to come out right.

Now we did say some time ago that conditional probabilities are just like ordinary probabilities, and whatever we do in probability theory can also be done in conditional universes. Talking about conditional probabilities. So since we have a notion of independence, then there should be also a notion of conditional independence. So independence was defined by the probability that A intersection B is equal to the probability of A times the probability of B.

What would be a reasonable definition of conditional independence? Conditional independence would mean that this same property could be true, but in a conditional universe where we are told that the certain event happens. So if we're told that the event C has happened, then we're transported in a conditional universe where the only thing that matters are conditional

probabilities. And this is just the same plain, previous definition of independence, but applied in a conditional universe.

So this is the definition of conditional independence. So it's independence, but with reference to the conditional probabilities. And intuitively it has, again, the same meaning, that in the conditional world, if I tell you that A occurred, then that doesn't change your beliefs about B.

So suppose you had a picture like this. And somebody told you that events A and B are independent unconditionally. Then somebody comes and tells you that event C actually has occurred, so we now live in this new universe. In this new universe, is the independence of A and B going to be preserved or not? Are A and B independent in this new universe?

The answer is no, because in the new universe, whatever is left of event A is this piece. Whatever is left of event B is this piece. And these two pieces are disjoint. So we are back in a situation of this kind. So in the conditional universe, A and B are disjoint. And therefore, generically, they're not going to be independent.

What's the moral of this example? Having independence in the original model does not imply independence in a conditional model.

The opposite is also possible. And let's illustrate by another example. So I have two coins, and both of them are badly biased. One coin is much biased in favor of heads. The other coin is much biased in favor of tails. So the probabilities being 90%.

Let's consider independent flips of coin A. This is the relevant model. This is a model of two independent flips of the first coin. There's going to be two flips, and each one has probability 0.9 of being heads. So that's a model that describes coin A. You can think of this as a conditional model which is a model of the coin flips conditioned on the fact that they have chosen coin A.

Alternatively we could be dealing with coin B In a conditional world where we chose coin B and flip it twice, this is the relevant model. The probability of two heads, for example, is the probability of heads the first time, heads the second time, and each one is 0.1.

Now I'm building this into a bigger experiment in which I first start by choosing one of the two coins at random. So I have these two coins. I blindly pick one of them. And then I start flipping them.

So the question now is, are the coin flips, or the coin tosses, are they independent of each other? If we just stay inside this sub-model here, are the coin flips independent? They are independent, because the probability of heads in the second toss is the same, 0.9, no matter what happened in the first toss. So the conditional probabilities of what happens in the second toss are not affected by the outcome of the first toss. So the second toss and the first toss are independent. So here we're just dealing with plain, independent coin flips.

Similarity the coin flips within this sub-model are also independent. Now the question is, if we look at the big model as just one probability model, instead of looking at the conditional sub-

models, are the coin flips independent of each other? Does the outcome of a few coin flips give you information about subsequent coin flips?

Well if I observe ten heads in a row-- So instead of two coin flips, now let's think of doing more of them so that the tree gets expanded.

So let's start with this. I don't know which coin it is. What's the probability that the 11th coin toss is going to be heads? There's complete symmetry here, so the answer could not be anything other than 1/2. So let's justify it, why is it 1/2?

Well, the probability that the 11th toss is heads, how can that outcome happen? It can happen in two ways. You can choose coin A, which happens with probability 1/2. And having chosen coin A, there's probability 0.9 that it results in that you get heads in the 11th toss. Or you can choose coin B. And if it's coin B when you flip it, there's probably 0.1 that you have heads. So the final answer is 1/2.

So each one of the coins is biased, but they're biased in different ways. If I don't know which coin it is, their two biases kind of cancel out, and the probability of obtaining heads is just in the middle, then it's 1/2.

Now if someone tells you that the first ten tosses were heads, is that going to change your beliefs about the 11th toss? Here's how a reasonable person would think about it.

If it's coin B the probability of obtaining 10 heads in a row is negligible. It's going to be 0.1 to the 10th. If it's coin A. The probability of 10 heads in a row is a more reasonable number. It's 0.9 to the 10th. So this event is a lot more likely to occur with coin A, rather than coin B.

The plausible explanation of having seen ten heads in a row is that I actually chose coin A. When you see ten heads in a row, you are pretty certain that it's coin A that we're dealing with. And once you're pretty certain that it's coin A that we're dealing with, what's the probability that the next toss is heads? It's going to be 0.9.

So essentially here I'm doing an inference calculation. Given this information, I'm making an inference about which coin I'm dealing with. I become pretty certain that it's coin A, and given that it's coin A, this probability is going to be 0.9. And I'm putting an approximate sign here, because the inference that I did is approximate. I'm pretty certain it's coin A. I'm not 100% certain that it's coin A.

But in any case what happens here is that the unconditional probability is different from the conditional probability. This information here makes me change my beliefs about the 11th toss. And this means that the 11th toss is dependent on the previous tosses. So the coin tosses have now become dependent. What is the physical link that causes this dependence? Well, the physical link is the choice of the coin. By choosing a particular coin, I'm introducing a pattern in the future coin tosses. And that pattern is what causes dependence.

OK, so I've been playing a little bit too loose with the language here, because we defined the concept of independence of two events. But here I have been referring to independent coin tosses, where I'm thinking about many coin tosses, like 10 or 11 of them.

So to be proper, I should have defined for you also the notion of independence of multiple events, not just two. We don't want to just say coin toss one is independent from coin toss two. We want to be able to say something like, these 10 then coin tosses are all independent of each other. Intuitively what that means should be the same thing-- that information about some of the coin tosses doesn't change your beliefs about the remaining coin tosses. How do we translate that into a mathematical definition?

Well, an ugly attempt would be to impose requirements such as this. Think of A1 being the event that the first flip was heads. A2 is the event of that the second flip was heads. A3, the third flip, was heads, and so on.

Here is an event whose occurrence is not determined by the first three coin flips. And here's an event whose occurrence or not is determined by the fifth and sixth coin flip. If we think physically that all those coin flips have nothing to do with each other, information about the fifth and sixth coin flip are not going to change what we expect from the first three. So the probability of this event, the conditional probability, should be the same as the unconditional probability. And we would like a relation of this kind to be true, no matter what kind of formula you write down, as long as the events that show up here are different from the events that show up there.

OK. That's sort of an ugly definition. The mathematical definition that actually does the job, and leads to all the formulas of this kind, is the following. We're going to say that the collection of events are independent if we can find the probability of their joint occurrence by just multiplying probabilities. And that will be true even if you look at sub-collections of these events.

Let's make that more precise. If we have three events, the definition tells us that the three events are independent if the following are true. Probability A1 and A2 and A3, you can calculate this probability by multiplying individual probabilities. But the same is true even if you take fewer events. Just a few indices out of the indices that we have available. So we also require $P(A_1 \text{ intersection } A_2)$ is $P(A_1) \times P(A_2)$. And similarly for the other possibilities of choosing the indices.

OK, so independence, mathematical definition, requires that calculating probabilities of any intersection of the events we have in our hands, that calculation can be done by just multiplying individual probabilities. And this has to apply to the case where we consider all of the events in our hands or just sub-collections of those events.

Now these relations just by themselves are called pairwise independence. So this relation, for example, tells us that A1 is independent from A2. This tells us that A2 is independent from A3. This will tell us that A1 is independent from A3. But independence of all the events together actually requires a little more. One more equality that has to do with all three events being considered at the same time.

And this extra equality is not redundant. It actually does make a difference. Independence and pairwise independence are different things. So let's illustrate the situation with an example. Suppose we have two coin flips. The coin tosses are independent, so the bias is $1/2$, so all possible outcomes have a probability of $1/2$ times $1/2$, which is $1/4$.

And let's consider now a bunch of different events. One event is that the first toss is heads. This is this blue set here. Another event is the second toss is heads. And this is this black event here.

OK. Are these two events independent? If you check it mathematically, yes. Probability of A is probability of B is $1/2$. Probability of A times probability of B is $1/4$, which is the same as the probability of A intersection B, which is this set. So we have just checked mathematically that A and B are independent.

Now let's consider a third event which is that the first and second toss give the same result. I'll use a different color. First and second toss to give the same result. This is the event that we obtain heads, heads or tails, tails. So this is the probability of C. What's the probability of C? Well, C is made up of two outcomes, each one of which has probability $1/4$, so the probability of C is $1/2$. What is the probability of C intersection A? C intersection A is just this one outcome, and has probability $1/4$.

What's the probability of A intersection B intersection C? The three events intersect just this outcome, so this probability is also $1/4$.

OK. What's the probability of C given A and B?

If A has occurred, and B has occurred, you are certain that this outcome here happened. If the first toss is H and the second toss is H, then you're certain of the first and second toss gave the same result. So the conditional probability of C given A and B is equal to 1.

So do we have independence in this example? We don't. C, that we obtain the same result in the first and the second toss, has probability $1/2$. Half of the possible outcomes give us two coin flips with the same result-- heads, heads or tails, tails. So the probability of C is $1/2$.

But if I tell you that the events A and B both occurred, then you're certain that C occurred. If I tell you that we had heads and heads, then you're certain the outcomes were the same. So the conditional probability is different from the unconditional probability. So by combining these two relations together, we get that the three events are not independent.

But are they pairwise independent? Is A independent from B? Yes, because probability of A times probability of B is $1/4$, which is probability of A intersection B. Is C independent from A? Well, the probability of C and A is $1/4$. The probability of C is $1/2$. The probability of A is $1/2$. So it checks. $1/4$ is equal to $1/2$ and $1/2$, so event C and event A are independent.

Knowing that the first toss was heads does not change your beliefs about whether the two tosses are going to have the same outcome or not. Knowing that the first was heads, well, the second is equally likely to be heads or tails. So event C has just the same probability, again, $1/2$, to occur.

To put it the opposite way, if I tell you that the two results were the same-- so it's either heads, heads or tails, tails-- what does that tell you about the first toss? Is it heads, or is it tails? Well, it doesn't tell you anything. It could be either over the two, so the probability of heads in the first toss is equal to $1/2$, and telling you C occurred does not change anything.

So this is an example that illustrates the case where we have three events in which we check that pairwise independence holds for any combination of two of these events. We have the probability of their intersection is equal to the product of their probabilities. On the other hand, the three events taken all together are not independent. A doesn't tell me anything useful, whether C is going to occur or not. B doesn't tell me anything useful. But if I tell you that both A and B occurred, the two of them together tell me something useful about C. Namely, they tell me that C certainly has occurred.

Very good. So independence is this somewhat subtle concept. Once you grasp the intuition of what it really means, then things perhaps fall in place. But it's a concept where it's easy to get some misunderstanding. So just take some time to digest.

So to lighten things up, I'm going to spend the remaining four minutes talking about the very nice, simple problem that involves conditional probabilities and the like. So here's the problem, formulated exactly as it shows up in various textbooks. And the formulation says the following.

Well, consider one of those anachronistic places where they still have kings or queens, and where actually boys take precedence over girls. So if there is a boy-- if the royal family has a boy, then he will become the king even if he has an older sister who might be the queen.

So we have one of those royal families. That royal family had two children, and we know that there is a king. There is a king, which means that at least one of the two children was a boy. Otherwise we wouldn't have a king. What is the probability that the king's sibling is female?

OK. I guess we need to make some assumptions about genetics. Let's assume that every child is a boy or a girl with probability $1/2$, and that different children, what they are is independent from what the other children were. So every childbirth is basically a coin flip.

OK, so if you take that, you say, well, the king is a child. His sibling is another child. Children are independent of each other. So the probability that the sibling is a girl is $1/2$. That's the naive answer. Now let's try to do it formally.

Let's set up a model of the experiment. The royal family had two children, as we're told, so there's four outcomes-- boy boy, boy girl, girl boy, and girl girl. Now, we are told that there is a king, which means what? This outcome here did not happen. It is not possible. There are three outcomes that remain possible. So this is our conditional sample space given that there is king.

What are the probabilities for the original model? Well with the model that we assume that every child is a boy or a girl independently with probability $1/2$, then the four outcomes would be equally likely, and they're like this. These are the original probabilities. But once we are told that

this outcome did not happen, because we have a king, then we are transported to the smaller sample space.

In this sample space, what's the probability that the sibling is a girl? Well the sibling is a girl in two out of the three outcomes. So the probability that the sibling is a girl is actually $2/3$. So that's supposed to be the right answer. Maybe a little counter-intuitive.

So you can play smart and say, oh I understand such problems better than you, here is a trick problem and here's why the answer is $2/3$. But actually I'm not fully justified in saying that the answer is $2/3$. I made lots of hidden assumptions when I put this model down, which I didn't yet state. So to reverse engineer this answer, let's actually think what's the probability model for which this would have been the right answer. And here's the probability model.

The royal family-- the royal parents decided to have exactly two children. They went and had them. It turned out that at least one was a boy and became a king. Under this scenario-- that they decide to have exactly two children-- then this is the big sample space. It turned out that one was a boy. That eliminates this outcome. And then this picture is correct and this is the right answer.

But there's hidden assumptions being there. How about if the royal family had followed the following strategy? We're going to have children until we get a boy, so that we get a king, and then we'll stop. OK, given they have two children, what's the probability that the sibling is a girl?

It's 1. The reason that they had two children was because the first was a girl, so they had to have a second. So assumptions about reproductive practices actually need to come in, and they're going to affect the decisions. Or, if it's one of those ancient kingdoms where a king would always make sure to strangle any of his brothers, then the probability that the sibling is a girl is actually 1 again, and so on.

So it means that one needs to be careful when you start with loosely worded problems to make sure exactly what it means and what assumptions you're making. All right, see you next week.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013 Transcript – Tutorial: Network Reliability

Previously, we learned the concept of independent experiments. In this exercise, we'll see how the seemingly simple idea of independence can help us understand the behavior of quite complex systems. In particular, we'll combine the concept of independence with the idea of divide and conquer, where we break a larger system into smaller components, and then use independent properties to glue them back together.

Now, let's take a look at the problem. We are given a network of connected components, and each component can be good with probability P or bad otherwise. All components are independent from each other. We say the system is operational if there exists a path connecting point A here to point B that goes through only the good components. And we'd like to understand, what is the probability that system is operational? Which we'll denote by P of A to B.

Although the problem might seem a little complicated at the beginning, it turns out only two structures really matter. So let's look at each of them. In the first structure, which we call the serial structure, we have a collection of k components, each one having probability P being good, connected one next to each other in a serial line.

Now, in this structure, in order for there to be a good path from A to B, every single one of the components must be working. So the probability of having a good path from A to B is simply P times P , so on and so, repeated k times, which is P raised to the k power. Know that the reason we can write the probability this way, in terms of this product, is because of the independence property.

Now, the second useful structure is parallel structure. Here again, we have k components one, two, through k , but this time they're connected in parallel to each other, namely they start from one point here and ends at another point here, and this holds for every single component. Now, for the parallel structure to work, namely for there to exist a good path from A to B, it's easy to see that as long as one of these components works the whole thing will work. So the probability of A to B is the probability that at least one of these components works. Or in the other word, the probability of the complement of the event where all components fail.

Now, if each component has probability P to be good, then the probability that all key components fail is $1 - P$ raised to the k th power. Again, having this expression means that we have used the property of independence, and that is probability of having a good parallel structure. Now, there's one more observation that will be useful for us. Just like how we define two components to be independent, we can also find two collections of components to be independent from each other.

For example, in this diagram, if we call the components between points C and E as collection two, and the components between E and B as collection three. Now, if we assume that each component in both collections-- they're completely independent from each other, then it's not

hard to see that collection two and three behave independently. And this will be very helpful in getting us the breakdown from complex networks to simpler elements.

Now, let's go back to the original problem of calculating the probability of having a good path from point big A to point big B in this diagram. Based on that argument of independent collections, we can first divide the whole network into three collections, as you see here, from A to C, C to E and E to B. Now, because they're independent and in a serial structure, as seen by the definition of a serial structure here, we see that the probability of A to B can be written as a probability of A to C multiplied by C to E, and finally, E to B.

Now, the probability of A to C is simply P because the collection contains only one element. And similarly, the probability of E to B is not that hard knowing the parallel structure here. We see that collection three has two components in parallel, so this probability will be given by 1 minus 1 minus P squared. And it remains to calculate just the probability of having a good path from point C to point E. To get a value for P C to E, we notice again, that this area can be treated as two components, C1 to E and C2 to E connected in parallel.

And using the parallel law we get this probability is 1 minus 1 minus P C1 to E multiplied by the 1 minus P C2 to E. Know that I'm using two different characters, C1 and C2, to denote the same node, which is C. This is simply for making it easier to analyze two branches where they actually do note the same node. Now P C1 to E is another serial connection of these three elements here with another component. So the first three elements are connected in parallel, and we know the probability of that being successful is 1 minus P3, and the last one is P.

And finally, P C2 to E. It's just a single element component with probability of successful being P. At this point, there is no longer any unknown variables, and we have indeed obtained exact values for all the quantities that we're interested in. So starting from this equation, we can plug in the values for P C2 to E, P C1 to E back here, and then further plug in P C to E back here. That will give us the final solution, which is given by the following somewhat complicated formula.

So in summary, in this problem, we learned how to use the independence property among different components to break down the entire fairly complex network into simple modular components, and use the law of serial and parallel connections to put the probabilities back together in common with the overall success probability of finding a path from A to B.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Tutorial 1
September 16/17, 2010

1. Let A and B be events such that $A \subset B$. Can A and B be independent?
2. An electrical system consists of identical components that are operational with probability p independently of other components. The components are connected in three subsystems, as shown in the figure. The system is operational if there is a path that starts at point A, ends at point B, and consists of operational components. This is the same as requiring that all three subsystems are operational. What are the probabilities that the three subsystems, as well as the entire system, are operational?

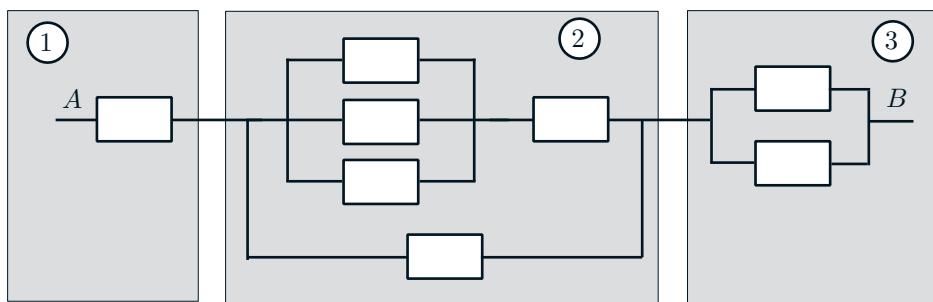


Figure 1: A system of identical components that consists of the three subsystems 1, 2, and 3. The system is operational if there is a path that starts at point A, ends at point B, and consists of operational components.

3. **The Chess Problem.** This year's Belmont chess champion is to be selected by the following procedure. Bo and Ci, the leading challengers, first play a two-game match. If one of them wins both games, he gets to play a two-game *second round* with Al, the current champion. Al retains his championship unless a second round is required and the challenger beats Al in both games. If Al wins the initial game of the second round, no more games are played.

Furthermore, we know the following:

- The probability that Bo will beat Ci in any particular game is 0.6.
- The probability that Al will beat Bo in any particular game is 0.5.
- The probability that Al will beat Ci in any particular game is 0.7.

Assume no tie games are possible and all games are independent.

- (a) Determine the apriori probabilities that
 - i. the second round will be required.
 - ii. Bo will win the first round.
 - iii. Al will retain his championship this year.
- (b) Given that the second round is required, determine the conditional probabilities that

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Fall 2010)

- i. Bo is the surviving challenger.
 - ii. Al retains his championship.
- (c) Given that the second round was required and that it comprised only one game, what is the conditional probability that it was Bo who won the first round?

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Problem Set 2
Due September 22, 2010

1. Most mornings, Victor checks the weather report before deciding whether to carry an umbrella. If the forecast is “rain,” the probability of actually having rain that day is 80%. On the other hand, if the forecast is “no rain,” the probability of it actually raining is equal to 10%. During fall and winter the forecast is “rain” 70% of the time and during summer and spring it is 20%.
 - (a) One day, Victor missed the forecast and it rained. What is the probability that the forecast was “rain” if it was during the winter? What is the probability that the forecast was “rain” if it was during the summer?
 - (b) The probability of Victor missing the morning forecast is equal to 0.2 on any day in the year. If he misses the forecast, Victor will flip a fair coin to decide whether to carry an umbrella. On any day of a given season he sees the forecast, if it says “rain” he will always carry an umbrella, and if it says “no rain,” he will not carry an umbrella. Are the events “Victor is carrying an umbrella,” and “The forecast is no rain” independent? Does your answer depend on the season?
 - (c) Victor is carrying an umbrella and it is not raining. What is the probability that he saw the forecast? Does it depend on the season?
2. You have a fair five-sided die. The sides of the die are numbered from 1 to 5. Each die roll is independent of all others, and all faces are equally likely to come out on top when the die is rolled. Suppose you roll the die twice.
 - (a) Let event A to be “the total of two rolls is 10”, event B be “at least one roll resulted in 5”, and event C be “at least one roll resulted in 1”.
 - i. Is event A independent of event B ?
 - ii. Is event A independent of event C ?
 - (b) Let event D be “the total of two rolls is 7”, event E be “the difference between the two roll outcomes is exactly 1”, and event F be “the second roll resulted in a higher number than the first roll”.
 - i. Are events E and F independent?
 - ii. Are events E and F independent given event D ?
3. The local widget factory is having a blowout widget sale. Everything must go, old and new. The factory has 500 old widgets, and 1500 new widgets in stock. The problem is that 15% of the old widgets are defective, and 5% of the new ones are defective as well. You can assume that widgets are selected at random when an order comes in. You are the first customer since the sale was announced.
 - (a) You flip a fair coin once to decide whether to buy old or new widgets. You order two widgets of the same type, chosen based on the outcome of the coin toss. What is the probability that they will both be defective?
 - (b) Given that both widgets turn out to be defective, what is the probability that they were old widgets?

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Fall 2010)

4. Oscar has lost his dog in either forest A (with a priori probability 0.4) or in forest B (with a priori probability 0.6).

On any given day, if the dog is in A and Oscar spends a day searching for it in A , the conditional probability that he will find the dog that day is 0.25. Similarly, if the dog is in B and Oscar spends a day looking for it there, the conditional probability that he will find the dog that day is 0.15.

The dog cannot go from one forest to the other. Oscar can search only in the daytime, and he can travel from one forest to the other only at night.

- (a) In which forest should Oscar look to maximize the probability he finds his dog on the first day of the search?
 - (b) Given that Oscar looked in A on the first day but didn't find his dog, what is the probability that the dog is in A ?
 - (c) If Oscar flips a fair coin to determine where to look on the first day and finds the dog on the first day, what is the probability that he looked in A ?
 - (d) If the dog is alive and not found by the N th day of the search, it will die that evening with probability $\frac{N}{N+2}$. Oscar has decided to look in A for the first two days. What is the probability that he will find a live dog for the first time on the second day?
5. In solving this problem, feel free to browse problems 43-45 in Chapter 1 of the text for ideas. If you need to, you may quote the results of these problems.

- (a) Suppose that A , B , and C are independent. Use the definition of independence to show that A and $B \cup C$ are independent.
- (b) Prove that if A_1, \dots, A_n are independent events, then

$$\mathbf{P}(A_1 \cup A_2 \cup \dots \cup A_n) = 1 - \prod_{i=1}^n (1 - \mathbf{P}(A_i)).$$

- G1[†]. Alice, Bob, and Caroll play a chess tournament. The first game is played between Alice and Bob. The player who sits out a given game plays next the winner of that game. The tournament ends when some player wins two successive games. Let a tournament history be the list of game winners, so for example *ACBAA* corresponds to the tournament where Alice won games 1, 4, and 5, Caroll won game 2, and Bob won game 3.

- (a) Provide a tree-based sequential description of a sample space where the outcomes are the possible tournament histories.
- (b) We are told that every possible tournament history that consists of k games has probability $1/2^k$, and that a tournament history consisting of an infinite number of games has zero probability. Demonstrate that this assignment of probabilities defines a legitimate probability law.
- (c) Assuming the probability law from part (b) to be correct, find the probability that the tournament lasts no more than 5 games, and the probability for each of Alice, Bob, and Caroll winning the tournament.

[†]Required for 6.431; optional for 6.041

MIT OpenCourseWare
<http://ocw.mit.edu>

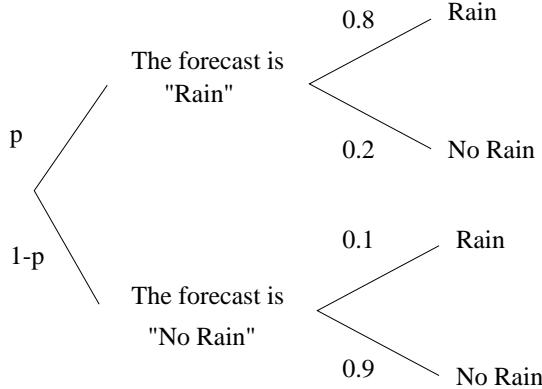
6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Problem Set 2: Solutions
Due September 22, 2010

1. (a) The tree representation during the winter can be drawn as the following:



Let A be the event that the forecast was “Rain,”
 let B be the event that it rained, and
 let p be the probability that the forecast says “Rain.” If it is in the winter, $p = 0.7$ and

$$\mathbf{P}(A | B) = \frac{\mathbf{P}(B | A)\mathbf{P}(A)}{\mathbf{P}(B)} = \frac{(0.8)(0.7)}{(0.8)(0.7) + (0.1)(0.3)} = \frac{56}{59}.$$

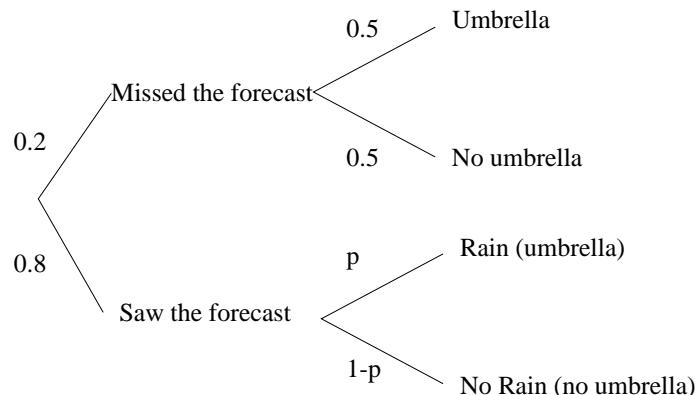
Similarly, if it is in the summer, $p = 0.2$ and

$$\mathbf{P}(A | B) = \frac{\mathbf{P}(B | A)\mathbf{P}(A)}{\mathbf{P}(B)} = \frac{(0.8)(0.2)}{(0.8)(0.2) + (0.1)(0.8)} = \frac{2}{3}.$$

- (b) Let C be the event that Victor is carrying an umbrella.

Let D be the event that the forecast is no rain.

The tree diagram in this case is:



$$\mathbf{P}(D) = 1 - p$$

$$\mathbf{P}(C) = (0.8)p + (0.2)(0.5) = 0.8p + 0.1$$

$$\mathbf{P}(C | D) = (0.8)(0) + (0.2)(0.5) = 0.1$$

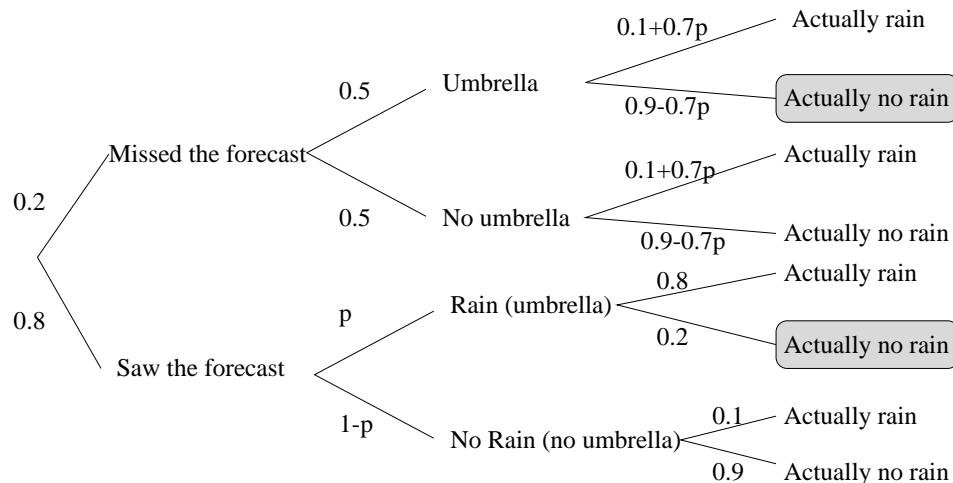
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
 (Fall 2010)

Therefore, $\mathbf{P}(C) = \mathbf{P}(C | D)$ if and only if $p = 0$. However, p can only be 0.7 or 0.2, which implies the events C and D can never be independent, and this result does not depend on the season.

- (c) Let us first find the probability of rain if Victor missed the forecast.

$$\mathbf{P}(\text{actually rains} | \text{missed forecast}) = (0.8)p + (0.1)(1 - p) = 0.1 + 0.7p.$$

Then, we can extend the tree in part (b) as follows:



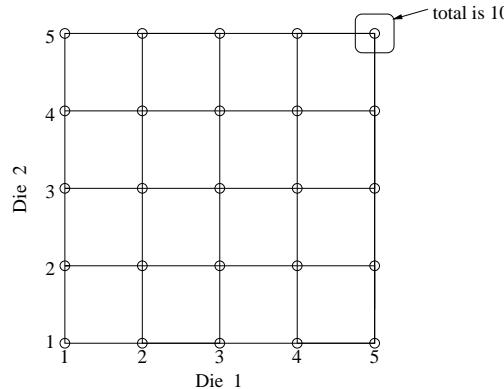
Therefore, given that Victor is carrying an umbrella and it is not raining, we are looking at the two shaded cases.

$$\mathbf{P}(\text{saw forecast} | \text{umbrella and not raining}) = \frac{(0.8)p(0.2)}{(0.8)p(0.2) + (0.2)(0.5)(0.9 - 0.7p)}$$

In fall and winter, $p = 0.7$, so the probability is $\frac{112}{153}$.

In summer and spring, $p = 0.2$, so the probability is $\frac{8}{27}$.

2. (a) i. No



Overall, there are 25 different outcomes in the sample space. For a total of 10, we should get a 5 on both rolls. Therefore $A \subset B$, and

$$\mathbf{P}(B|A) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(A)} = \frac{\mathbf{P}(A)}{\mathbf{P}(A)} = 1$$

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
 (Fall 2010)

We observe that to get at least one 5 showing, we can have 5 on the first roll, 5 on the second roll, or 5 on both rolls, which corresponds to 9 distinct outcomes in the sample space. Therefore

$$\mathbf{P}(B) = \frac{9}{25} \neq \mathbf{P}(B|A)$$

- ii. **No** Given event A , we know that both roll outcomes must be 5. Therefore, we could not have event C occur, which would require at least one 1 showing. Formally, there are 9 outcomes in C , and

$$\mathbf{P}(C) = \frac{9}{25}$$

But

$$\mathbf{P}(C|A) = 0 \neq \mathbf{P}(C)$$

- (b) i. **No** Out of the total 25 outcomes, 5 outcomes correspond to equal numbers in the two rolls. In half of the remaining 20 outcomes, the second number is higher than the first one. In the other half, the first number is higher than the second. Therefore,

$$\mathbf{P}(F) = \frac{10}{25}$$

There are eight outcomes that belong to event E :

$$E = \{(1, 2), (2, 3), (3, 4), (4, 5), (2, 1), (3, 2), (4, 3), (5, 4)\}.$$

To find $\mathbf{P}(F|E)$, we need to compute the proportion of outcomes in E for which the second number is higher than the first one:

$$\mathbf{P}(F|E) = \frac{1}{2} \neq \mathbf{P}(F)$$

- ii. **Yes** Conditioning on event D reduces the sample space to just four outcomes

$$\{(2, 5), (3, 4), (4, 3), (5, 2)\}$$

which are all equally likely. It is easy to see that

$$\mathbf{P}(E|D) = \frac{2}{4} = \frac{1}{2}, \quad \mathbf{P}(F|D) = \frac{2}{4} = \frac{1}{2}, \quad \mathbf{P}(E \cap F|D) = \frac{1}{4} = \mathbf{P}(E|D)\mathbf{P}(F|D)$$

3. (a) Suppose we choose old widgets. Before we choose any widgets, there are $500 \cdot 0.15 = 75$ defective old widgets. The probability that we choose two defective widgets is

$$\begin{aligned} \mathbf{P}(\text{two defective|old}) &= \mathbf{P}(\text{first is defective|old}) \cdot \mathbf{P}(\text{second is defective|first is defective, old}) \\ &= \frac{75}{500} \frac{74}{499} = 0.02224 \end{aligned}$$

Now let's consider the new widgets. Before we choose any widgets, there are $1500 \cdot 0.05 = 75$ defective old widgets. Similar to the calculations above,

$$\begin{aligned} \mathbf{P}(\text{two defective|new}) &= \mathbf{P}(\text{first is defective|new}) \cdot \mathbf{P}(\text{second is defective|first is defective, new}) \\ &= \frac{75}{1500} \frac{74}{1499} = 0.002568 \end{aligned}$$

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
 (Fall 2010)

By the total probability law,

$$\begin{aligned}\mathbf{P}(\text{two defective}) &= \mathbf{P}(\text{old}) \cdot \mathbf{P}(\text{two defective}|\text{old}) \\ &\quad + \mathbf{P}(\text{new}) \cdot \mathbf{P}(\text{two defective}|\text{new}) \\ &= \frac{1}{2} \cdot 0.02224 + \frac{1}{2} \cdot 0.002568 = 0.01240.\end{aligned}$$

Note that this number is very close to what we would get if we ignored the effects of removing one defective widget before choosing the second widget:

$$\begin{aligned}\mathbf{P}(\text{two defective}) &= \mathbf{P}(\text{old}) \cdot \mathbf{P}(\text{two defective}|\text{old}) \\ &\quad + \mathbf{P}(\text{new}) \cdot \mathbf{P}(\text{two defective}|\text{new}) \\ &\approx \frac{1}{2} \cdot 0.15^2 + \frac{1}{2} \cdot 0.05^2 = 0.0125.\end{aligned}$$

(b) Using Bayes' rule,

$$\begin{aligned}\mathbf{P}(\text{old}|\text{two defective}) &= \frac{\mathbf{P}(\text{old}) \cdot \mathbf{P}(\text{two defective}|\text{old})}{\mathbf{P}(\text{old}) \cdot \mathbf{P}(\text{two defective}|\text{old}) + \mathbf{P}(\text{new}) \cdot \mathbf{P}(\text{two defective}|\text{new})} \\ &= \frac{\frac{1}{2} \cdot 0.02224}{\frac{1}{2} \cdot 0.02224 + \frac{1}{2} \cdot 0.002568} = 0.8965\end{aligned}$$

4. (a)

$$\mathbf{P}(\text{find in A and in A}) = \mathbf{P}(\text{in A}) \cdot \mathbf{P}(\text{find in A}|\text{in A}) = 0.4 \cdot 0.25 = 0.1$$

$$\mathbf{P}(\text{find in B and in B}) = \mathbf{P}(\text{in B}) \cdot \mathbf{P}(\text{find in B}|\text{in B}) = 0.6 \cdot 0.15 = 0.09$$

Oscar should search in Forest A first.

(b) Using Bayes' Rule,

$$\begin{aligned}\mathbf{P}(\text{in A}|\text{not find in A}) &= \frac{\mathbf{P}(\text{not find in A}|\text{in A}) \cdot \mathbf{P}(\text{in A})}{\mathbf{P}(\text{not find in A}|\text{in A}) \cdot \mathbf{P}(\text{in A}) + \mathbf{P}(\text{not find in A}|\text{in B}) \cdot \mathbf{P}(\text{in B})} \\ &= \frac{(0.75) \cdot (0.4)}{(0.4) \cdot (0.75) + (1) \cdot (0.6)} = \frac{1}{3}\end{aligned}$$

(c) Again, using Bayes' Rule,

$$\begin{aligned}\mathbf{P}(\text{looked in A}| \text{find dog}) &= \frac{\mathbf{P}(\text{find dog}|\text{looked in A}) \cdot \mathbf{P}(\text{looked in A})}{\mathbf{P}(\text{find dog})} \\ &= \frac{(0.25) \cdot (0.4) \cdot (0.5)}{(0.25) \cdot (0.4) \cdot (0.5) + (0.15) \cdot (0.6) \cdot (0.5)} = \frac{10}{19}\end{aligned}$$

(d) In order for Oscar to find the dog, it must be in Forest A, not found on the first day, alive, and found on the second day. Note that this calculation requires conditional independence of not finding the dog on different days and the dog staying alive.

$$\begin{aligned}\mathbf{P}(\text{find live dog in A day 2}) &= \mathbf{P}(\text{in A}) \cdot \mathbf{P}(\text{not find in A day 1}|\text{in A}) \\ &\quad \cdot \mathbf{P}(\text{alive day 2}) \cdot \mathbf{P}(\text{find day 2}|\text{in A}) \\ &= 0.4 \cdot 0.75 \cdot \left(1 - \frac{1}{3}\right) \cdot 0.25 = 0.05\end{aligned}$$

5. (a) We proceed as follows:

$$\begin{aligned}
 \mathbf{P}(A \cap (B \cup C)) &= \mathbf{P}((A \cap B) \cup (A \cap C)) \\
 &= \mathbf{P}(A \cap B) + \mathbf{P}(A \cap C) - \mathbf{P}(A \cap B \cap C) \\
 &\stackrel{*}{=} \mathbf{P}(A)\mathbf{P}(B) + \mathbf{P}(A)\mathbf{P}(C) - \mathbf{P}(A)\mathbf{P}(B)\mathbf{P}(C) \\
 &= \mathbf{P}(A)[\mathbf{P}(B) + \mathbf{P}(C) - \mathbf{P}(B)\mathbf{P}(C)] \\
 &= \mathbf{P}(A)\mathbf{P}(B \cup C),
 \end{aligned}$$

where the equality marked with * follows from the independence of A , B , and C .

(b) Proof 1: If A and B are independent, then A^c and B^c are also independent (see Problem 1.43, page 63 for the proof).

For any two independent events U and V , DeMorgan's Law implies

$$\begin{aligned}
 \mathbf{P}(U \cup V) &= \mathbf{P}((U^c \cap V^c)^c) = 1 - \mathbf{P}(U^c \cap V^c) = 1 - \mathbf{P}(U^c) \cdot \mathbf{P}(V^c) \\
 &= 1 - (1 - \mathbf{P}(U))(1 - \mathbf{P}(V)).
 \end{aligned}$$

We proceed to prove the statement by induction. Letting $U = A_1$ and $V = A_2$, the base case is proven above. Now we assume that the result holds for any n and show that it holds for $n + 1$. For independent $\{A_1, \dots, A_n, A_{n+1}\}$, let $B = \cup_{i=1}^n A_i$. It is easy to show that B and A_{n+1} are independent. Therefore,

$$\begin{aligned}
 \mathbf{P}(A_1 \cup A_2 \cup \dots \cup A_{n+1}) &= 1 - (1 - \mathbf{P}(B)) \cdot (1 - \mathbf{P}(A_{n+1})) \\
 &= 1 - \prod_{i=1}^{n+1} (1 - \mathbf{P}(A_i)),
 \end{aligned}$$

which completes the proof.

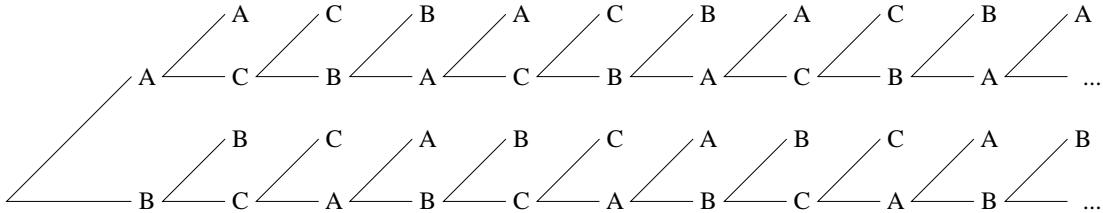
Proof 2: Alternatively, we can use the version of the DeMorgan's Law for n events:

$$\begin{aligned}
 \mathbf{P}(A_1 \cup A_2 \cup \dots \cup A_n) &= \mathbf{P}((A_1^c \cap A_2^c \cap \dots \cap A_n^c)^c) \\
 &= 1 - \mathbf{P}(A_1^c \cap A_2^c \cap \dots \cap A_n^c).
 \end{aligned}$$

But we know that $A_1^c, A_2^c, \dots, A_n^c$ are independent. Therefore

$$\begin{aligned}
 \mathbf{P}(A_1 \cup A_2 \cup \dots \cup A_n) &= 1 - \mathbf{P}(A_1^c)\mathbf{P}(A_2^c) \dots \mathbf{P}(A_n^c) \\
 &= 1 - \prod_{i=1}^n (1 - \mathbf{P}(A_i)).
 \end{aligned}$$

G1[†]. (a) The figure below describes the sample space via an infinite tree. The leaves of this tree are exactly all *finite* tournament histories; in addition, the two infinite paths represent the two *infinite* tournament histories that are possible. Note that the winner of the first game is either Alice or Bob; from then on, the winner of a game is either the winner of the previous game (in which case we have reached a leaf and the tournament has ended) or the player that sat out the previous game. The outcomes of the sample space correspond to the finite histories (which are identified with the leaves of the tree) and the two infinite histories: ACBACB... and BCABCA...



- (b) The probability of an event is $1/2^k$ times the number of finite histories contained in the event. The probability of the event consisting of one or both infinite histories is 0. We have to show that this probability law satisfies the three probability axioms. It clearly satisfies nonnegativity and additivity. To check *normalization*, we have to verify that the probabilities of all tournament histories sum up to 1.

Start by noticing that two of the histories are infinite and have probability 0. Each one of the remaining histories has some finite length $k \geq 2$ (and hence is represented by one of the two leaves of the tree of the figure above at depth k) and probability $1/2^k$. Hence, summing all probabilities we get

$$2 \cdot 0 + \sum_{k=2}^{\infty} 2 \cdot \frac{1}{2^k} = \sum_{k=2}^{\infty} \frac{1}{2^{k-1}} = \sum_{k=0}^{\infty} \frac{1}{2^{k+1}} = \frac{1}{2} \sum_{k=0}^{\infty} \frac{1}{2^k} = \frac{1}{2} \frac{1}{1 - 1/2} = 1.$$

- (c) The probability that exactly 2 games will be played is the sum of the probabilities of the two leaves at depth 2; that is,

$$P(\text{exactly 2 games}) = \frac{1}{2^2} + \frac{1}{2^2} = \frac{1}{2}.$$

Similarly, the probability that exactly i games will be played, for $i = 3, 4, 5$, is

$$\begin{aligned} P(\text{exactly 3 games}) &= \frac{1}{2^3} + \frac{1}{2^3} = \frac{1}{4}, \\ P(\text{exactly 4 games}) &= \frac{1}{2^4} + \frac{1}{2^4} = \frac{1}{8}, \\ P(\text{exactly 5 games}) &= \frac{1}{2^5} + \frac{1}{2^5} = \frac{1}{16}. \end{aligned}$$

Hence, the probability that the tournament lasts no more than 5 games is

$$P(\text{at most 5 games}) = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} = \frac{15}{16}.$$

Hence, it's pretty probable that the tournament will last at most that much.

The probability that Alice wins the tournament is the sum of the probabilities of the leaves of the tree that are labeled "A"; that is,

$$\left(\frac{1}{2^2} + \frac{1}{2^5} + \frac{1}{2^8} + \dots \right) + \left(\frac{1}{2^4} + \frac{1}{2^7} + \frac{1}{2^{10}} + \dots \right),$$

where the first summation includes all leaves from the upper part of the tree, while the second one takes care of the leaves on the lower part. Calculating, we have

$$\frac{1}{4} \left(1 + \frac{1}{2^3} + \frac{1}{2^6} + \dots \right) + \frac{1}{16} \left(1 + \frac{1}{2^3} + \frac{1}{2^6} + \dots \right) = \frac{5}{16} \sum_{j=0}^{\infty} \frac{1}{8^j} = \frac{5}{16} \frac{1}{1 - 1/8} = \frac{5}{14}.$$

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Fall 2010)

By symmetry (note the correspondence between the histories where Alice wins and the histories where Bob does), Bob's probability of winning is $\frac{5}{14}$, as well. Then, since the outcomes where nobody wins (these are the two infinite tournament histories) have total probability 0, Carol wins with probability $1 - \frac{5}{14} - \frac{5}{14} = \frac{4}{14}$. Hence, by not participating in the first game, Carol enters the tournament with a disadvantage.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

LECTURE 3

- **Readings:** Section 1.5
- Review
- Independence of two events
- Independence of a collection of events

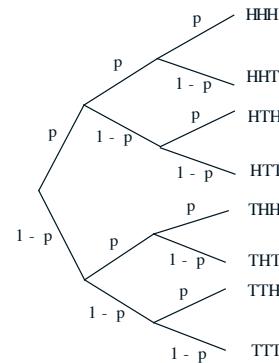
Review

$$P(A | B) = \frac{P(A \cap B)}{P(B)}, \quad \text{assuming } P(B) > 0$$

- Multiplication rule:
$$P(A \cap B) = P(B) \cdot P(A | B) = P(A) \cdot P(B | A)$$
- Total probability theorem:
$$P(B) = P(A)P(B | A) + P(A^c)P(B | A^c)$$
- Bayes rule:
$$P(A_i | B) = \frac{P(A_i)P(B | A_i)}{P(B)}$$

Models based on conditional probabilities

- 3 tosses of a biased coin:
 $P(H) = p, P(T) = 1 - p$



$$P(THT) =$$

$$P(1 \text{ head}) =$$

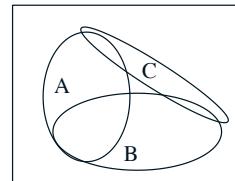
$$P(\text{first toss is H} | 1 \text{ head}) =$$

Independence of two events

- “Defn:” $P(B | A) = P(B)$
 - “occurrence of A provides no information about B 's occurrence”
- Recall that $P(A \cap B) = P(A) \cdot P(B | A)$
- **Defn:** $P(A \cap B) = P(A) \cdot P(B)$
- Symmetric with respect to A and B
 - applies even if $P(A) = 0$
 - implies $P(A | B) = P(A)$

Conditioning may affect independence

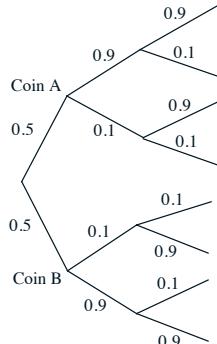
- Conditional independence, given C , is defined as independence under probability law $P(\cdot | C)$
- Assume A and B are independent



- If we are told that C occurred, are A and B independent?

Conditioning may affect independence

- Two unfair coins, A and B :
 $P(H | \text{coin } A) = 0.9$, $P(H | \text{coin } B) = 0.1$
choose either coin with equal probability



- Once we know it is coin A , are tosses independent?
- If we do not know which coin it is, are tosses independent?
 - Compare:
 $P(\text{toss } 11 = H)$
 $P(\text{toss } 11 = H | \text{first 10 tosses are heads})$

Independence of a collection of events

- Intuitive definition:
Information on some of the events tells us nothing about probabilities related to the remaining events

– E.g.:

$$P(A_1 \cap (A_2^c \cup A_3) | A_5 \cap A_6^c) = P(A_1 \cap (A_2^c \cup A_3))$$

- Mathematical definition:
Events A_1, A_2, \dots, A_n
are called **independent** if:

$$P(A_i \cap A_j \cap \dots \cap A_q) = P(A_i)P(A_j) \cdots P(A_q)$$

for any distinct indices i, j, \dots, q ,
(chosen from $\{1, \dots, n\}$)

Independence vs. pairwise independence

- Two independent fair coin tosses
 - A : First toss is H
 - B : Second toss is H
 - $P(A) = P(B) = 1/2$

HH	HT
TH	TT

- C : First and second toss give same result
 - $P(C) =$
 - $P(C \cap A) =$
 - $P(A \cap B \cap C) =$
 - $P(C | A \cap B) =$
- Pairwise independence **does not** imply independence

The king's sibling

- The king comes from a family of two children. What is the probability that his sibling is female?

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Recitation 3: September 16, 2010

1. Example 1.20, page 37 in the text.

Consider two independent fair coin tosses, in which all four possible outcomes are equally likely. Let

$$\begin{aligned} H_1 &= \{\text{1st toss is a head}\}, \\ H_2 &= \{\text{2nd toss is a head}\}, \\ D &= \{\text{the two tosses produced different results}\}. \end{aligned}$$

- (a) Are the events H_1 and H_2 (unconditionally) independent?
 (b) Given event D has occurred, are the events H_1 and H_2 (conditionally) independent?
2. Imagine a drunk tightrope walker, in the middle of a really long tightrope, who manages to keep his balance, but takes a step forward with probability p and takes a step back with probability $(1 - p)$.
 - What is the probability that after two steps the tightrope walker will be at the same place on the rope?
 - What is the probability that after three steps, the tightrope walker will be one step forward from where he began?
 - Given that after three steps he has managed to move ahead one step, what is the probability that the first step he took was a step forward?
3. Problem 1.31, page 60 in the text.

Communication through a noisy channel. A binary (0 or 1) message transmitted through a noisy communication channel is received incorrectly with probability ϵ_0 and ϵ_1 , respectively (see the figure). Errors in different symbol transmissions are independent. The channel source transmits a 0 with probability p and transmits a 1 with probability $1 - p$.

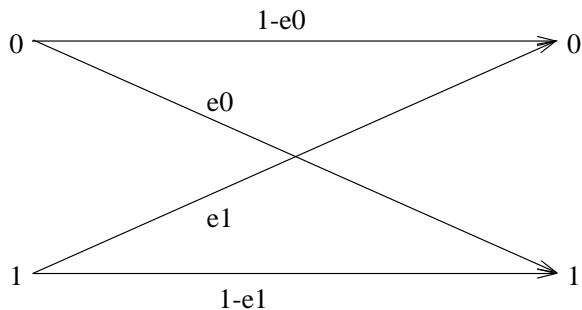


Figure 1: Error probabilities in a binary communication channel.

- What is the probability that a randomly chosen symbol is received correctly?
- Suppose that the string of symbols 1011 is transmitted. What is the probability that all the symbols in the string are received correctly?

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Fall 2010)

- (c) In an effort to improve reliability, each symbol is transmitted three times and the received symbol is decoded by majority rule. In other words, a 0 (or 1) is transmitted as 000 (or 111, respectively), and it is decoded at the receiver as a 0 (or 1) if and only if the received three-symbol string contains at least two 0s (or 1s, respectively). What is the probability that a transmitted 0 is correctly decoded?
- (d) Suppose that the scheme of part (c) is used. What is the probability that a 0 was transmitted given that the received string is 101?
4. (a) Can an event A be independent of itself?
(b) Problem 1.43(a) on page 63 in text.
Let A and B be independent events. Use the definition of independence to prove that the events A and B^c are independent.
- (c) Problem 1.44 on page 64 in text.
Let A , B , and C be independent events, with $\mathbf{P}(C) > 0$. Prove that A and B are conditionally independent of C .

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Recitation 3: Solutions
September 16, 2010

1. See the textbook, Example 1.20, page 37.
2. (a) In order to wind up in the same place after two steps, the tightrope walker can either step forwards, then backwards, or vice versa. Therefore the required probability is:

$$2 \cdot p \cdot (1 - p).$$

- (b) The probability that after three steps he will be one step ahead of his starting point is the probability that out of 3 steps in total, 2 of them are forwards, and one is backwards. This equals:

$$3 \cdot p^2 \cdot (1 - p).$$

- (c) Given that out of his three steps only one is backwards, the sample space for the experiment is:

$$\{(F, F, B); (F, B, F); (B, F, F)\}$$

where F denotes a step forwards, and B a step backwards. Each of these sample points is equally likely, therefore the probability that his first step is a step forward is $\frac{2}{3}$.

3. See the textbook, Problem 1.31, page 60.
4. (a) A is independent of itself if and only if $\mathbf{P}(A \cap A) = \mathbf{P}(A)\mathbf{P}(A)$. Since $A \cap A = A$ then A must satisfy $\mathbf{P}(A) = (\mathbf{P}(A))^2$. Therefore, A is independent of itself if and only if $\mathbf{P}(A) = 1$ or $\mathbf{P}(A) = 0$.
- (b) See solution to Problem 1.43(a) in text on pages 63-64.
- (c) See solution to Problem 1.44 in text on page 64.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

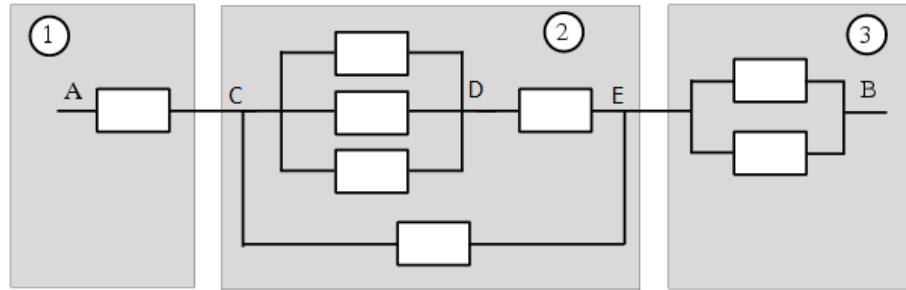
Tutorial 1 Solutions
September 16/17, 2010

1. If $A \subset B$, then $\mathbf{P}(B \cap A) = \mathbf{P}(A)$. But we know that in order for A and B to be independent, $\mathbf{P}(B \cap A) = \mathbf{P}(A)\mathbf{P}(B)$. Therefore, A and B are independent if and only if $\mathbf{P}(B) = 1$ or $\mathbf{P}(A) = 0$. This could happen, for example, if B is the universe or if A is empty.
2. This problem is similar in nature to Example 1.24, page 40. In order to compute the success probability of individual sub-systems, we make use of the following two properties, derived in that example:
 - If a *serial* sub-system contains m components with success probabilities $p_1, p_2 \dots p_m$, then the probability of success of the entire sub-system is given by

$$\mathbf{P}(\text{whole system succeeds}) = p_1 p_2 p_3 \dots p_m$$

- If a *parallel* sub-system contains m components with success probabilities $p_1, p_2 \dots p_m$, then the probability of success of the entire sub-system is given by

$$\mathbf{P}(\text{whole system succeeds}) = 1 - (1 - p_1)(1 - p_2)(1 - p_3) \dots (1 - p_m)$$



Let $\mathbf{P}(X \rightarrow Y)$ denote the probability of a successful connection between node X and Y . Then,

$$\begin{aligned}\mathbf{P}(A \rightarrow B) &= \mathbf{P}(A \rightarrow C)\mathbf{P}(C \rightarrow E)\mathbf{P}(E \rightarrow B) \quad (\text{since they are in series}) \\ \mathbf{P}(A \rightarrow C) &= p \\ \mathbf{P}(C \rightarrow E) &= 1 - (1 - p)(1 - \mathbf{P}(C \rightarrow D)\mathbf{P}(D \rightarrow E)) \\ \mathbf{P}(E \rightarrow B) &= 1 - (1 - p)^2\end{aligned}$$

The probabilities $\mathbf{P}(C \rightarrow D)$, $\mathbf{P}(D \rightarrow E)$ can be similarly computed as

$$\begin{aligned}\mathbf{P}(C \rightarrow D) &= 1 - (1 - p)^3 \\ \mathbf{P}(D \rightarrow E) &= p\end{aligned}$$

The probability of success of the entire system can be obtained by substituting the subsystem success probabilities:

$$\mathbf{P}(A \rightarrow B) = p(1 - (1 - p)(1 - (1 - (1 - p)^3)p))(1 - (1 - p)^2).$$

3. The Chess Problem.

- (a) i. $\mathbf{P}(\text{2nd Rnd Req}) = (0.6)^2 + (0.4)^2 = 0.52$
ii. $\mathbf{P}(\text{Bo Wins 1st Rnd}) = (0.6)^2 = 0.36$
iii. $\mathbf{P}(\text{Al Champ}) = 1 - \mathbf{P}(\text{Bo Champ}) - \mathbf{P}(\text{Ci Champ})$
 $= 1 - (0.6)^2 * (0.5)^2 - (0.4)^2 * (0.3)^2 = 0.8956$
- (b) i. $\mathbf{P}(\text{Bo Challenger}|\text{2nd Rnd Req}) = \frac{(0.6)^2}{0.52} = \frac{0.36}{0.52} = 0.6923$
ii. $\mathbf{P}(\text{Al Champ}|\text{2nd Rnd Req})$
 $= \mathbf{P}(\text{Al Champ}|\text{Bo Challenger, 2nd Rnd Req}) \times \mathbf{P}(\text{Bo Challenger}|\text{2nd Rnd Req})$
 $+ \mathbf{P}(\text{Al Champ}|\text{Ci Challenger, 2nd Rnd Req}) \times \mathbf{P}(\text{Ci Challenger}|\text{2nd Rnd Req})$
 $= (1 - (0.5)^2) \times 0.6923 + (1 - (0.3)^2) \times 0.3077$
 $= 0.7992$
- (c) $\mathbf{P}((\text{Bo Challenger})|\{(2\text{nd Rnd Req}) \cap (\text{One Game})\}) = \frac{(0.6)^2 * (0.5)}{(0.6)^2 * (0.5) + (0.4)^2 * (0.7)}$
 $= \frac{(0.6)^2 * (0.5)}{0.2920} = 0.6164$

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013

Transcript – Recitation: Hypergeometric Probabilities

In this problem, we're given an urn with n balls in it, out of which m balls are red balls. To visualize it, we can draw a box that represents the set of all n balls. Somewhere in the middle or somewhere else we have a cut, such that to the left we have all the red balls (there are m), and non-red balls. Let's for now call it black balls. That is n minus m .

Now, from this box, we are to draw k balls, and we'd like to know the probability that i out of those k balls are red balls. For the rest of the problem, we'll refer to this probability as p_r , where r stands for the red balls. So from this picture, we know that we're going to draw a subset of the balls, such that i of them are red, and the remaining k minus i are black. And we'll like to know what is the probability that this event would occur.

To start, we define our sample space, Ω , as the set of all ways to draw k balls out of n balls. We found a simple counting argument -- we know that size of our sample space has n -choose- k , which is the total number of ways to draw k balls out of n balls.

Next, we'd like to know how many of those samples correspond to the event that we're interested in. In particular, we would like to know c , which is equal to the number of ways to get i red balls after we draw the k balls. To do so, we'll break c into a product of two numbers -- let's call it a times b -- where a is the total number of ways to select i red balls out of m red balls. So the number of ways to get i out of m red balls.

Going back to the picture, this corresponds to the total number of ways to get these balls. And similarly, we define b as the total number of ways to get the remaining k minus i balls out of the set n minus m black balls. This corresponds to the total number of ways to select the subset right here in the right side of the box.

Now as you can see, once we have a and b , we multiply them together, and this yields the total number of ways to get i red balls. To compute what these numbers are, we see that a is equal to m -choose- i number of ways to get i red balls, and b is n minus m , the total number of black balls, choose k minus i , the balls that are not red within those k balls.

Now putting everything back, we have p_r , the probability we set out to compute, is equal to c , the size of the event, divided by the size of the entire sample space. From the previous calculations, we know that c is equal to a times b , which is then equal to m -choose- i times $(n$ minus $m)$ -choose- $(k$ minus $i)$. And on the denominator, we have the entire sample space is a size n -choose- k . And that completes our derivation.

Now let's look at a numerical example of this problem. Here, let's say we have a deck of 52 cards. And we draw a box with n equals 52, out of which we know that there are 4 aces. So we'll call these the left side of the box, which is we have m equals 4 aces. Now if we were to draw

seven cards-- call it k equal to 7-- and we'd like to know what is the probability that out of the 7 cards, we have 3 aces.

Using the notation we did earlier, if we were to draw a circle representing the seven cards, we want to know what is the probability that we have 3 aces in the left side of the box and 4 non-aces for the remainder of the deck. In particular, we'll call i equal to 3. So by this point, we've cast the problem of drawing cards from the deck in the same way as we did earlier of drawing balls from an urn.

And from the expression right here, which we computed earlier, we can readily compute the probability of having 3 aces. In particular, we just have to substitute into the expression right here the value of m equal to 4, n equal to 52, k equal to 7, finally, i equal to 3. So we have 4-choose-3 times n minus m , in this case would be 48, choose k minus i , will be 4, and on the denominator, we have 52 total number of cards, choosing 7 cards. That gives us [the] numerical answer [for] the probability of getting 3 aces when we draw 7 cards.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013 Transcript – Lecture 4

The following content is provided under a Creative Commons license. Your support will help MIT OpenCourseWare continue to offer high quality educational resources for free. To make a donation, or view additional materials from hundreds of MIT courses, visit MIT OpenCourseWare at ocw.mit.edu.

PROFESSOR: OK. So today's lecture will be on the subject of counting. So counting, I guess, is a pretty simple affair conceptually, but it's a topic that can also get to be pretty tricky. The reason we're going to talk about counting is that there's a lot of probability problems whose solution actually reduces to successfully counting the cardinalities of various sets.

So we're going to see the basic, simplest methods that one can use to count systematically in various situations. So in contrast to previous lectures, we're not going to introduce any significant new concepts of a probabilistic nature. We're just going to use the probability tools that we already know. And we're going to apply them in situations where there's also some counting involved.

Now, today we're going to just touch the surface of this subject. There's a whole field of mathematics called combinatorics who are people who actually spend their whole lives counting more and more complicated sets. We were not going to get anywhere close to the full complexity of the field, but we'll get just enough tools that allow us to address problems of the type that one encounters in most common situations.

So the basic idea, the basic principle is something that we've already discussed. So counting methods apply in situations where we have probabilistic experiments with a finite number of outcomes and where every outcome-- every possible outcome-- has the same probability of occurring.

So we have our sample space, Ω , and it's got a bunch of discrete points in there. And the cardinality of the set Ω is some capital N . So, in particular, we assume that the sample points are equally likely, which means that every element of the sample space has the same probability equal to 1 over N .

And then we are interested in a subset of the sample space, call it A . And that subset consists of a number of elements. Let the cardinality of that subset be equal to little n . And then to find the probability of that set, all we need to do is to add the probabilities of the individual elements. There's little n elements, and each one has probability one over capital N . And that's the answer.

So this means that to solve problems in this context, all that we need to be able to do is to figure out the number capital N and to figure out the number little n . Now, if somebody gives you a set by just giving you a list and gives you another set, again, giving you a list, it's easy to count there element. You just count how much there is on the list.

But sometimes the sets are described in some more implicit way, and we may have to do a little bit more work. There's various tricks that are involved in counting properly. And the most common one is to-- when you consider a set of possible outcomes, to describe the construction of those possible outcomes through a sequential process.

So think of a probabilistic experiment that involves a number of stages, and in each one of the stages there's a number of possible choices that there may be. The overall experiment consists of carrying out all the stages to the end. And the number of points in the sample space is how many final outcomes there can be in this multi-stage experiment.

So in this picture we have an experiment in which of the first stage we have four choices. In the second stage, no matter what happened in the first stage, the way this is drawn we have three choices. No matter whether we ended up here, there, or there, we have three choices in the second stage.

And then there's a third stage and at least in this picture, no matter what happened in the first two stages, in the third stage we're going to have two possible choices. So how many leaves are there at the end of this tree? That's simple. It's just the product of these three numbers. The number of possible leaves that we have out there is 4 times 3 times 2. Number of choices at each stage gets multiplied, and that gives us the number of overall choices.

So this is the general rule, the general trick that we are going to use over and over. So let's apply it to some very simple problems as a warm up. How many license plates can you make if you're allowed to use three letters and then followed by four digits? At least if you're dealing with the English alphabet, you have 26 choices for the first letter. Then you have 26 choices for the second letter. And then 26 choices for the third letter.

And then we start the digits. We have 10 choices for the first digit, 10 choices for the second digit, 10 choices for the third, 10 choices for the last one. Let's make it a little more complicated, suppose that we're interested in license plates where no letter can be repeated and no digit can be repeated. So you have to use different letters, different digits. How many license plates can you make?

OK, let's choose the first letter, and we have 26 choices. Now, I'm ready to choose my second letter, how many choices do I have? I have 25, because I already used one letter. I have the 25 remaining letters to choose from. For the next letter, how many choices? Well, I used up two of my letters, so I only have 24 available.

And then we start with the digits, 10 choices for the first digit, 9 choices for the second, 8 for the third, 7 for the last one. All right. So, now, let's bring some symbols in a related problem. You are given a set that consists of n elements and you're supposed to take those n elements and put them in a sequence. That is to order them. Any possible ordering of those elements is called a permutation.

So for example, if we have the set 1, 2, 3, 4, a possible permutation is the list 2, 3, 4, 1. That's one possible permutation. And there's lots of possible permutations, of course, the question is

how many are there. OK, let's think about building this permutation by choosing one at a time. Which of these elements goes into each one of these slots? How many choices for the number that goes into the first slot or the elements?

Well, we can choose any one of the available elements, so we have n choices. Let's say this element goes here, having used up that element, we're left with n minus 1 elements and we can pick any one of these and bring it into the second slot. So here we have n choices, here we're going to have n minus 1 choices, then how many we put there will have n minus 2 choices. And you go down until the end.

What happens at this point when you are to pick the last element? Well, you've used n minus of them, there's only one left in your bag. You're forced to use that one. So the last stage, you're going to have only one choice. So, basically, the number of possible permutations is the product of all integers from n down to one, or from one up to n . And there's a symbol that we use for this number, it's called n factorial.

So n factorial is the number of permutations of n objects. The number of ways that you can order n objects that are given to you. Now, a different equation. We have n elements. Let's say the elements are 1, 1, 2, up to n . And it's a set. And we want to create a subset. How many possible subsets are there?

So speaking of subsets means looking at each one of the elements and deciding whether you're going to put it in to subsets or not. For example, I could choose to put 1 in, but 2 I'm not putting it in, 3 I'm not putting it in, 4 I'm putting it, and so on. So that's how you create a subset. You look at each one of the elements and you say, OK, I'm going to put it in the subset, or I'm not going to put it.

So think of these as consisting of stages. At each stage you look at one element, and you make a binary decision. Do I put it in the subset, or not? So therefore, how many subsets are there? Well, I have two choices for the first element. Am I going to put in the subset, or not? I have two choices for the next element, and so on.

For each one of the elements, we have two choices. So the overall number of choices is 2 to the power n . So, conclusion-- the number of subsets, often n element set, is 2 to the n . So in particular, if we take n equal to 1, let's check that our answer makes sense. If we have n equal to one, how many subsets does it have?

So we're dealing with a set of just one. What are the subsets? One subset is this one. Do we have other subsets of the one element set? Yes, we have the empty set. That's the second one. These are the two possible subsets of this particular set. So 2 subsets when n is equal to 1, that checks the answer.

All right. OK, so having gone so far, we can do our first example now. So we are given a die and we're going to roll it 6 times. OK, let's make some assumptions about the rolls. Let's assume that the rolls are independent, and that the die is also fair.

So this means that the probability of any particular outcome of the die rolls-- for example, so we have 6 rolls, one particular outcome could be 3,3,1,6,5. So that's one possible outcome. What's the probability of this outcome? There's probability $1/6$ that this happens, $1/6$ that this happens, $1/6$ that this happens, and so on. So the probability that the outcome is this is $1/6$ to the sixth.

What did I use to come up with this answer? I used independence, so I multiplied the probability of the first roll gives me a 2, times the probability that the second roll gives me a 3, and so on. And then I used the assumption that the die is fair, so that the probability of 2 is $1/6$, the probability of 3 is $1/6$, and so on.

So if I were to spell it out, it's the probability that we get the 2 in the first roll, times the probability of 3 in the second roll, times the probability of the 5 in the last roll. So by independence, I can multiply probabilities. And because the die is fair, each one of these numbers is $1/6$ to the sixth.

And so the same calculation would apply no matter what numbers I would put in here. So all possible outcomes are equally likely. Let's start with this. So since all possible outcomes are equally likely to find an answer to a probability question, if we're dealing with some particular event, so the event is that all rolls give different numbers. That's our event A. And our sample space is some set capital omega. We know that the answer is going to be the cardinality of the set A, divided by the cardinality of the set omega.

So let's deal with the easy one first. How many elements are there in the sample space? How many possible outcomes are there when you roll a dice 6 times?

You have 6 choices for the first roll. You have 6 choices for the second roll and so on. So the overall number of outcomes is going to be 6 to the sixth. So number of elements in the sample space is 6 to the sixth power.

And I guess this checks with this. We have 6 to the sixth outcomes, each one has this much probability, so the overall probability is equal to one. Right? So the probability of an individual outcome is one over how many possible outcomes we have, which is this. All right.

So how about the numerator? We are interested in outcomes in which the numbers that we get are all different. So what is an outcome in which the numbers are all different? So the die has 6 faces. We roll it 6 times. We're going to get 6 different numbers. This means that we're going to exhaust all the possible numbers, but they can appear in any possible sequence.

So an outcome that makes this event happen is a list of the numbers from 1 to 6, but arranged in some arbitrary order. So the possible outcomes that make event A happen are just the permutations of the numbers from 1 to 6.

One possible outcome that makes our events to happen-- it would be this. Here we have 6 possible numbers, but any other list of this kind in which none of the numbers is repeated would also do. So number of outcomes that make the event happen is the number of permutations of 6

elements. So it's 6 factorial. And so the final answer is going to be 6 factorial divided by 6 to the sixth.

All right, so that's a typical way that's one solves problems of this kind. We know how to count certain things. For example, here we knew how to count permutations, and we used our knowledge to count the elements of the set that we need to deal with.

So now let's get to a slightly more difficult problem. We're given once more a set with n elements. We already know how many subsets that set has, but now we would be interested in subsets that have exactly k elements in them. So we start with our big set that has n elements, and we want to construct a subset that has k elements.

Out of those n I'm going to choose k and put them in there. In how many ways can I do this? More concrete way of thinking about this problem-- you have n people in some group and you want to form a committee by picking people from that group, and you want to form a committee with k people. Where k is a given number. For example, a 5 person committee. How many 5 person committees are possible if you're starting with 100 people?

So that's what we want to count. How many k element subsets are there? We don't yet know the answer, but let's give a name to it. And the name is going to be this particular symbol, which we read as n choose k . Out of n elements, we want to choose k of them.

OK. That may be a little tricky. So what we're going to do is to instead figure out a somewhat easier problem, which is going to be-- in how many ways can I pick k out of these people and puts them in a particular order? So how many possible ordered lists can I make that consist of k people? By ordered, I mean that we take those k people and we say this is the first person in the community. That's the second person in the committee. That's the third person in the committee and so on.

So in how many ways can we do this? Out of these n , we want to choose just k of them and put them in slots. One after the other. So this is pretty much like the license plate problem we solved just a little earlier.

So we have n choices for who we put as the top person in the community. We can pick anyone and have them be the first person. Then I'm going to choose the second person in the committee. I've used up 1 person. So I'm going to have n minus 1 choices here.

And now, at this stage I've used up 2 people, so I have n minus 2 choices here. And this keeps going on. Well, what is going to be the last number? Is it's n minus k ? Well, not really. I'm starting subtracting numbers after the second one, so by the end I will have subtracted k minus 1. So that's how many choices I will have for the last person.

So this is the number of ways-- the product of these numbers there gives me the number of ways that I can create ordered lists consisting of k people out of the n that we started with. Now, you can do a little bit of algebra and check that this expression here is the same as that expression.

Why is this? This factorial has all the products from 1 up to n. This factorial has all the products from 1 up to n minus k. So you get cancellations. And what's left is all the products starting from the next number after here, which is this particular number.

So the number of possible ways of creating such ordered lists is n factorial divided by n minus k factorial. Now, a different way that I could make an ordered list-- instead of picking the people one at a time, I could first choose my k people who are going to be in the committee, and then put them in order. And tell them out of these k, you are the first, you are the second, you are the third.

Starting with this k people, in how many ways can I order them? That's the number of permutations. Starting with a set with k objects, in how many ways can I put them in a specific order? How many specific orders are there? That's basically the question. In how many ways can I permute these k people and arrange them.

So the number of ways that you can do this step is k factorial. So in how many ways can I start with a set with n elements, go through this process, and end up with a sorted list with k elements? By the rule that-- when we have stages, the total number of stages is how many choices we had in the first stage, times how many choices we had in the second stage. The number of ways that this process can happen is this times that.

This is a different way that that process could happen. And the number of possible of ways is this number. No matter which way we carry out that process, in the end we have the possible ways of arranging k people out of the n that we started with.

So the final answer that we get when we count should be either this, or this times that. Both are equally valid ways of counting, so both should give us the same answer. So we get this equality here. So these two expressions corresponds to two different ways of constructing ordered lists of k people starting with n people initially.

And now that we have this relation, we can send the k factorial to the denominator. And that tells us what that number, n choose k, is going to be. So this formula-- it's written here in red, because you're going to see it a zillion times until the end of the semester-- they are called the binomial coefficients. And they tell us the number of possible ways that we can create a k element subset, starting with a set that has n elements.

It's always good to do a sanity check to formulas by considering extreme cases. So let's take the case where k is equal to n. What's the right answer in this case? How many n elements subsets are there out of an element set?

Well, your subset needs to include every one. You don't have any choices. There's only one choice. It's the set itself. So the answer should be equal to 1. That's the number of n element subsets, starting with a set with n elements. Let's see if the formula gives us the right answer.

We have n factorial divided by k, which is n in our case-- n factorial. And then n minus k is 0 factorial. So if our formula is correct, we should have this equality. And what's the way to make

that correct? Well, it depends what kind of meaning do we give to this symbol? How do we define zero factorial?

I guess in some ways it's arbitrary. We're going to define it in a way that makes this formula right. So the definition that we will be using is that whenever you have 0 factorial, it's going to stand for the number 1. So let's check that this is also correct, at the other extreme case.

If we let k equal to 0, what does the formula give us? It gives us, again, n factorial divided by 0 factorial times n factorial. According to our convention, this again is equal to 1. So there is one subset of our set that we started with that has zero elements. Which subset is it? It's the empty set.

So the empty set is the single subset of the set that we started with that happens to have exactly zero elements. So the formula checks in this extreme case as well. So we're comfortable using it. Now these factorials and these coefficients are really messy algebraic objects.

There's lots of beautiful identities that they satisfy, which you can prove algebraically sometimes by using induction and having cancellations happen all over the place. But it's really messy. Sometimes you can bypass those calculations by being clever and using your understanding of what these coefficients stand for.

So here's a typical example. What is the sum of those binomial coefficients? I fix n , and sum over all possible cases. So if you're an algebra genius, you're going to take this expression here, plug it in here, and then start doing algebra furiously. And half an hour later, you may get the right answer.

But now let's try to be clever. What does this really do? What does that formula count? We're considering k element subsets. That's this number. And we're considering the number of k element subsets for different choices of k .

The first term in this sum counts how many 0-element subsets we have. The next term in this sum counts how many 1-element subsets we have. The next term counts how many 2-element subsets we have. So in the end, what have we counted?

We've counted the total number of subsets. We've considered all possible cardinalities. We've counted the number of subsets of size k . We've considered all possible sizes k . The overall count is going to be the total number of subsets. And we know what this is. A couple of slides ago, we discussed that this number is equal to 2 to the n .

So, nice, clean and simple answer, which is easy to guess once you give an interpretation to the algebraic expression that you have in front of you. All right. So let's move again to sort of an example in which those binomial coefficients are going to show up.

So here's the setting-- n independent coin tosses, and each coin toss has a probability, P , of resulting in heads. So this is our probabilistic experiment. Suppose we do 6 tosses. What's the probability that we get this particular sequence of outcomes?

Because of independence, we can multiply probability. So it's going to be the probability that the first toss results in heads, times the probability that the second toss results in tails, times the probability that the third one results in tails, times probability of heads, times probability of heads, times probability of heads, which is just P to the fourth times $(1 - P)$ squared.

So that's the probability of this particular sequence. How about a different sequence? If I had 4 tails and 2 heads, but in a different order-- let's say if we considered this particular outcome-- would the answer be different?

We would still have P , times P , times P , times P , times $(1 - P)$, times $(1 - P)$. We would get again, the same answer. So what you observe from just this example is that, more generally, the probability of obtaining a particular sequence of heads and tails is P to a power, equal to the number of heads. So here we had 4 heads. So there's P to the fourth showing up. And then $(1 - P)$ to the power number of tails.

So every k head sequence-- every outcome in which we have exactly k heads, has the same probability, which is going to be P to the k , $(1 - p)$, to the $(n - k)$. This is the probability of any particular sequence that has exactly k heads. So that's the probability of a particular sequence with k heads.

So now let's ask the question, what is the probability that my experiment results in exactly k heads, but in some arbitrary order? So the heads could show up anywhere. So there's a number of different ways that this can happen. What's the overall probability that this event takes place?

So the probability of an event taking place is the sum of the probabilities of all the individual ways that the event can occur. So it's the sum of the probabilities of all the outcomes that make the event happen. The different ways that we can obtain k heads are the number of different sequences that contain exactly k heads.

We just figured out that any sequence with exactly k heads has this probability. So to do this summation, we just need to take the common probability of each individual k head sequence, times how many terms we have in this sum.

So what we're left to do now is to figure out how many k head sequences are there. How many outcomes are there in which we have exactly k heads. OK. So what are the ways that I can describe to you a sequence with k heads?

I can take my n slots that corresponds to the different tosses. I'm interested in particular sequences that have exactly k heads. So what I need to do is to choose k slots and assign heads to them. So to specify a sequence that has exactly k heads is the same thing as drawing this picture and telling you which are the k slots that happened to have heads.

So I need to choose out of those n slots, k of them, and assign them heads. In how many ways can I choose this k slots? Well, it's the question of starting with a set of n slots and choosing k slots out of the n available.

So the number of k head sequences is the same as the number of k element subsets of the set of slots that we started with, which are the n slots 1 up to n . We know what that number is. We counted, before, the number of k element subsets, starting with a set with n elements. And we gave a symbol to that number, which is that thing, n choose k . So this is the final answer that we obtain.

So these are the so-called binomial probabilities. And they gave us the probabilities for different numbers of heads starting with a fair coin that's being tossed a number of times. This formula is correct, of course, for reasonable values of k , meaning its correct for k equals 0, 1, up to n .

If k is bigger than n , what's the probability of k heads? If k is bigger than n , there's no way to obtain k heads, so that probability is, of course, zero. So these probabilities only makes sense for the numbers k that are possible, given that we have n tosses.

And now a question similar to the one we had in the previous slide. If I write down this summation-- even worse algebra than the one in the previous slide-- what do you think this number will turn out to be? It should be 1 because this is the probability of obtaining k heads.

When we do the summation, what we're doing is we're considering the probability of 0 heads, plus the probability of 1 head, plus the probability of 2 heads, plus the probability of n heads. We've exhausted all the possibilities in our experiment. So the overall probability, when you exhaust all possibilities, must be equal to 1. So that's yet another beautiful formula that evaluates into something really simple. And if you tried to prove this identity algebraically, of course, you would have to suffer quite a bit.

So now armed with the binomial probabilities, we can do the harder problems. So let's take the same experiment again. We flip a coin independently 10 times. So these 10 tosses are independent. We flip it 10 times. We don't see the result, but somebody comes and tells us, you know, there were exactly 3 heads in the 10 tosses that you had. OK? So a certain event happened.

And now you're asked to find the probability of another event, which is that the first 2 tosses were heads. Let's call that event A. OK. So are we in the setting of discrete uniform probability laws? When we toss a coin multiple times, is it the case that all outcomes are equally likely? All sequences are equally likely?

That's the case if you have a fair coin-- that all sequences are equally likely. But if your coin is not fair, of course, heads/heads is going to have a different probability than tails/tails. If your coin is biased towards heads, then heads/heads is going to be more likely.

So we're not quite in the uniform setting. Our overall sample space, Ω , does not have equally likely elements. Do we care about that? Not necessarily. All the action now happens inside the event B that we are told has occurred.

So we have our big sample space, Ω . Elements of that sample space are not equally likely. We are told that a certain event B occurred. And inside that event B , we're asked to find the conditional probability that A has also occurred.

Now here's the lucky thing, inside the event B , all outcomes are equally likely. The outcomes inside B are the sequences of 10 tosses that have exactly 3 heads. Every 3-head sequence has this probability. So the elements of B are equally likely with each other.

Once we condition on the event B having occurred, what happens to the probabilities of the different outcomes inside here? Well, conditional probability laws keep the same proportions as the unconditional ones. The elements of B were equally likely when we started, so they're equally likely once we are told that B has occurred.

So to do with this problem, we need to just transport us to this smaller universe and think about what's happening in that little universe. In that little universe, all elements of B are equally likely. So to find the probability of some subset of that set, we only need to count the cardinality of B , and count the cardinality of A . So let's do that.

Number of outcomes in B -- in how many ways can we get 3 heads out of 10 tosses? That's the number we considered before, and it's 10 choose 3. This is the number of 3-head sequences when you have 10 tosses.

Now let's look at the event A . The event A is that the first 2 tosses were heads, but we're living now inside this universe B . Given that B occurred, how many elements does A have in there? In how many ways can A happen inside the B universe.

If you're told that the first 2 were heads-- sorry. So out of the outcomes in B that have 3 heads, how many start with heads/heads? Well, if it starts with heads/heads, then the only uncertainty is the location of the third head.

So we started with heads/heads, we're going to have three heads, the question is, where is that third head going to be. It has eight possibilities. So slot 1 is heads, slot 2 is heads, the third heads can be anywhere else. So there's 8 possibilities for where the third head is going to be.

OK. So what we have counted here is really the cardinality of $A \cap B$, which is out of the elements in B , how many of them make A happen, divided by the cardinality of B . And that gives us the answer, which is going to be 10 choose 3, divided by 8.

And I should probably redraw a little bit of the picture that they have here. The set A is not necessarily contained in B . It could also have stuff outside B . So the event that the first 2 tosses are heads can happen with a total of 3 heads, but it can also happen with a different total number of heads.

But once we are transported inside the set B , what we need to count is just this part of A . It's $A \cap B$ and compare it with the total number of elements in the set B . Did I write it the opposite way? Yes. So this is 8 over 10 choose 3.

OK. So we're going to close with a more difficult problem now. OK. This business of n choose k has to do with starting with a set and picking a subset of k elements. Another way of thinking of that is that we start with a set with n elements and you choose a subset that has k , which means that there's n minus k that are left.

Picking a subset is the same as partitioning our set into two pieces. Now let's generalize this question and start counting partitions in general. Somebody gives you a set that has n elements. Somebody gives you also certain numbers-- n_1, n_2, n_3 , let's say, n_4 , where these numbers add up to n . And you're asked to partition this set into four subsets where each one of the subsets has this particular cardinality.

So you're asking to cut it into four pieces, each one having the prescribed cardinality. In how many ways can we do this partitioning? n choose k was the answer when we partitioned in two pieces, what's the answer more generally? For a concrete example of a partition, you have your 52 card deck and you deal, as in bridge, by giving 13 cards to each one of the players.

Assuming that the dealing is done fairly and with a well shuffled deck of cards, every particular partition of the 52 cards into four hands, that is four subsets of 13 each, should be equally likely. So we take the 52 cards and we partition them into subsets of 13, 13, 13, and 13. And we assume that all possible partitions, all possible ways of dealing the cards are equally likely. So we are again in a setting where we can use counting, because all the possible outcomes are equally likely.

So an outcome of the experiment is the hands that each player ends up getting. And when you get the cards in your hands, it doesn't matter in which order that you got them. It only matters what cards you have on you. So it only matters which subset of the cards you got.

All right. So what's the cardinality of the sample space in this experiment? So let's do it for the concrete numbers that we have for the problem of partitioning 52 cards. So think of dealing as follows-- you shuffle the deck perfectly, and then you take the top 13 cards and give them to one person. In how many possible hands are there for that person?

Out of the 52 cards, I choose 13 at random and give them to the first person. Having done that, what happens next? I'm left with 39 cards. And out of those 39 cards, I pick 13 of them and give them to the second person. Now I'm left with 26 cards.

Out of those 26, I choose 13, give them to the third person. And for the last person there isn't really any choice. Out of the 13, I have to give that person all 13. And that number is just equal to 1. So we don't care about it.

All right. So next thing you do is to write down the formulas for these numbers. So, for example, here you would have 52 factorial, divided by 13 factorial, times 39 factorial, and you continue. And then there are nice cancellations that happen.

This 39 factorial is going to cancel the 39 factorial that comes from there, and so on. After you do the cancellations and all the algebra, you're left with this particular answer, which is the

number of possible partitions of 52 cards into four players where each player gets exactly 13 hands.

If you were to generalize this formula to the setting that we have here, the more general formula is-- you have n factorial, where n is the number of objects that you are distributing, divided by the product of the factorials of the-- OK, here I'm doing it for the case where we split it into four sets. So that would be the answer when we partition a set into four subsets of prescribed cardinalities.

And you can guess how that formula would generalize if you want to split it into five sets or six sets. OK. So far we just figured out the size of the sample space. Now we need to look at our event, which is the event that each player gets an ace, let's call that event A. In how many ways can that event happen? How many possible hands are there in which every player has exactly one ace?

So I need to think about the sequential process by which I distribute the cards so that everybody gets exactly one ace, and then try to think in how many ways can that sequential process happen. So one way of making sure that everybody gets exactly one ace is the following-- I take the four aces and I distribute them randomly to the four players, but making sure that each one gets exactly one ace. In how many ways can that happen?

I take the ace of spades and I send it to a random person out of the four. So there's 4 choices for this. Then I'm left with 3 aces to distribute. That person already gotten an ace. I take the next ace, and I give it to one of the 3 people remaining. So there's 3 choices for how to do that.

And then for the next ace, there's 2 people who have not yet gotten an ace, and they give it randomly to one of them. So these are the possible ways of distributing for the 4 aces, so that each person gets exactly one. It's actually the same as this problem.

Starting with a set of four things, in how many ways can I partition them into four subsets where the first set has one element, the second has one element, the third one has another element, and so on. So it agrees with that formula by giving us 4 factorial.

OK. So there are different ways of distributing the aces. And then there's different ways of distributing the remaining 48 cards. How many ways are there? Well, I have 48 cards that I'm going to distribute to four players by giving 12 cards to each one. It's exactly the same question as the one we had here, except that now it's 48 cards, 12 to each person. And that gives us this particular count.

So putting all that together gives us the different ways that we can distribute the cards to the four players so that each one gets exactly one ace. The number of possible ways is going to be this four factorial, coming from here, times this number-- this gives us the number of ways that the event of interest can happen-- and then the denominator is the cardinality of our sample space, which is this number.

So this looks like a horrible mess. It turns out that this expression does simplify to something really, really simple. And if you look at the textbook for this problem, you will see an alternative derivation that gives you a short cut to the same numerical answer. All right. So that basically concludes chapter one. From next time we're going to consider introducing random variables and make the subject even more interesting.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013

Transcript – Recitation: Rooks on a Chessboard

Today, we're going to do a fun problem called rooks on a chessboard. And rooks on a chessboard is a problem that's going to test your ability on counting. So hopefully by now in class, you've learned a few tricks to approach counting problems. You've learned about permutations, you've learned about k-permutations, you've learned about combinations, and you've learned about partitions.

And historically for students that we've taught in the past and many people, counting can be a tricky topic. So this is just one drill problem to help you get those skills under your belt. So what does the rooks on a chessboard problem ask you? Well, you're given an 8-by-8 chessboard, which I've tried to draw here. It's not very symmetrical. Sorry about that.

And you're told that you have eight rooks. I'm sure most of you guys are familiar with chess. But if any of you aren't, chess is a sophisticated board game. And one of the types of pieces you have in this game is called a rook. And in this particular problem, there are eight rooks.

And your job is to place all eight rooks onto this 8-by-8 chessboard. Now, you're told in the problem statement that all placements of rooks are equally likely. And you are tasked with finding the probability that you get a safe arrangement. So that is to say, you place your eight rooks on the board. What is the probability that the way you placed them is safe?

So what do I mean by "safe"? Well, if you're familiar with the way chess works, so if you place a rook here, it can move vertically or it can move horizontally. Those are the only two legal positions. So if you place a rook here and you have another piece here, then this is not a safe arrangement, because the rook can move this way and kill you.

Similarly, if you have a rook here and another piece here, the rook can move horizontally and kill you that way. So two rooks on this board are only safe from each other if they are neither in the same column nor in the same row. And that's going to be key for us to solve this problem.

So let's see-- where did my marker go? I've been talking a lot, and I haven't really been writing anything. So our job is again, to find the probability that you get a safe arrangement. So I'm just going to do "arrange" for short.

Now, I talked about this previously, and you guys have heard it in lecture. Hopefully you remember something called the discrete uniform law. So the discrete uniform law is applicable when your sample space is discrete and all outcomes are equally likely. So let's do a quick check here.

What is our sample space for this problem? Well, a logical choice would be that the set of all possible outcomes is the set of all possible spatial arrangements of rooks. And hopefully it's clear

to you that that is discrete. And the problem statement furthermore gives us that they're equally likely.

So the discrete uniform law is in fact applicable in our setting. So I'm going to go ahead and write what this means. So when your sample space is discrete and all outcomes are equally likely, then you can compute the probability of any event, A, simply by counting the number of outcomes in A and then dividing it by the total number of outcomes in your sample space. So here we just have to find the number of total safe arrangements and then divide it by the total number of arrangements.

So again, as you've seen in other problems, the discrete uniform law is really nice, because you reduce the problem of computing probabilities to the problem of counting. And so here's where we're going to exercise those counting skills, as I promised earlier. Now, I would like to start with computing the denominator, or the total number of arrangements, because I think it's a slightly easier computation.

So we don't care about the arrangements being safe. We just care about how many possible arrangements are there. Now, again, we have eight rooks, and we need to place all of them. And we have this 8-by-8 board. So pretty quickly, you guys could probably tell me that the total number of squares is 64, because this is just 8 times 8.

Now, I like to approach problems sequentially. That sort of really helps me think clearly about them. So I want you to imagine a sequential process during which we place each rook one at a time. So pick a rook. The chessboard is currently empty. So how many squares can you place that rook in? Well, nobody's on the board.

You can place it in 64 spots. So for the first rook that you pick, there are 64 spots. Now, once you place this rook, you need to place the second rook, because again, we're not done until all eight are placed. So how many possible spots are left.

Well, I claim that there are 63, because one rule of chess is that if you put a piece in a particular square, you can no longer put anything else on that square. You can't put two or more things. So the first rook is occupying one spot, so there's only 63 spots left.

So the second rook has 63 spots that it could go in. Similarly, the third rook has 62 spots. Hopefully you see the pattern. You can continue this down. And remember, we have to place all eight rooks.

So you could do it out yourself or just do the simple math. You'll figure out that the eighth rook only has 57 spots that it could be in. So this is a good start. We've sort of figured out if we sequentially place each rook, how many options do we have. But we haven't combined these numbers in any useful way yet. We haven't counted the number of total arrangements.

And this may already be obvious to some, but it wasn't obvious to me when I was first learning this material, so I want to go through this slowly. You have probably heard in lecture by now about the counting principle. And what the counting principle tells you is that whenever you

have a process that is done in stages and in each stage, you have a particular number of choices, to get the total number of choices available at the end of the process, you simply multiply the number of choices at each stage.

This might be clear to you, again, simply from the statement, for some of you. But for others, it might still not be clear. So let's just take a simple example. Forget about the rook problem for a second.

Let's say you're at a deli, and you want to make a sandwich. And to make a sandwich, you need a choice of bread and you need a choice of meat. So we have a sandwich-building process, and there's two stages. First, you have to pick the bread, and then you have to pick the meat. So let's say for the choice of bread, you can choose wheat or rye.

So again, you can always use a little decision tree-- wheat or rye. And then let's say that for the meats, you have three options. You have ham, turkey, and salami. So you can have ham, turkey, or salami-- ham, turkey, or salami. How many total possible sandwiches can you make?

Well, six. And I got to that by 2 times 3. And hopefully this makes sense for you, because there's two options in the first stage. Freeze an option. Given this choice, there's three options at the second stage. But you have to also realize that for every other option you have at the first stage, you have to add an additional three options for the second stage.

And this is the definition of multiplication. If you add three two times, you know that's 3 times 2. So if you extrapolate this example to a larger, more general picture, you will have derived for yourself the counting principle. And we're going to use the counting principle here to determine what the total number of arrangements are.

So we have a sequential process, because we're placing the first rook and then the second rook, et cetera. So at the first stage, we have 64 choices. At the second stage, we have 63 choices. At the third stage, we have 62 choices, et cetera.

And so I'm just multiplying these numbers together, because the counting principle says I can do this. So my claim is that this product is equal to the total number of arrangements. And we could stop here, but I'm going to actually write this in a more useful way.

You guys should have been introduced to the factorial function. So you can express this equivalently as 64 factorial divided by 56 factorial. And this is not necessary for your problem solution, but sometimes it's helpful to express these types of products in factorials, because you can see cancellations more easily. So if it's OK with everybody, I'm going to erase this work to give myself more room.

So we'll just put our answer for the denominator up here, and then we're going to get started on the numerator. So for the numerator, thanks to the discrete uniform law, we only need to count the number of safe arrangements. But this is a little bit more tricky, because now, we have to apply our definition of what "safe" means. But we're going to use the same higher-level strategy, which is realizing that we can place rooks sequentially.

So we can think of it as a sequential process. And then if we figure out how many choices you have in each stage that sort of maintain the "safeness" of the setup, then you can use the counting principle to multiply all those numbers together and get your answer. So we have to place eight rooks.

Starting the same way we did last time, how many spots are there for the first rook that are safe? Nobody is on the board yet, so nobody can harm the first rook we put down. So I claim that it's just our total of 64. Now, let's see what happens. Let's pick a random square in here. Let's say we put our first rook here.

Now, I claim a bunch of spots get invalidated because of the rules of chess. So before, I told you a rook can kill anything in the same column or in the same row. So you can't put a rook here, because they'll kill each other, and you can't put a rook here. So by extension, you can see that everything in the column and the row that I'm highlighting in blue, it's no longer an option.

You can't place a rook in there. Otherwise, we will have violated our "safety" principle. So where can our second rook go? Well, our second rook can go in any of the blank spots, any of the spots that are not highlighted by blue. And let's stare at this a little bit.

Imagine that you were to take scissors to your chessboard and cut along this line and this line and this line and this line. So you essentially sawed off this cross that we created. Then you would have four free-floating chessboard pieces-- this one, this one, this one, and this one. So this is a 3-by-4 piece, this is 3-by-3, this is 4-by-3, and this is 4-by-4.

Well, because you cut this part out, you can now slide those pieces back together. And hopefully you can convince yourself that that would leave you with a 7-by-7 chessboard. And you can see that the dimensions match up here.

So essentially, the second rook can be placed anywhere in the remaining 7-by-7 chessboard. And of course, there are 49 spots in a 7-by-7 chessboard. So you get 49. So let's do this experiment again. Let me rewrite the reduced 7-by-7 chessboard. You're going to have to forgive me if the lines are not perfect-- one, two, three, four, five, six, seven; one, two, three, four, five, six, seven.

Yep, I did that right. And then we have one, two, three, four, five, six, seven. That's not too bad for my first attempt. So again, how did I get this chessboard from this one? Well, I took scissors and I cut off of the blue strips, and then I just merged the remaining four pieces. So now, I'm placing my second rook.

So I know that I can place my second rook in any of these squares, and it'll be safe from this rook. Of course, in reality, you wouldn't really cut up your chessboard. I'm just using this as a visual aid to help you guys see why there are 49 spots. Another way you could see 49 spots is literally just by counting all the white squares, but I think it takes time to count 49 squares.

And this is a faster way of seeing it. So you can put your second rook anywhere here. Let's actually put in the corner, because the corner is a nice case. If you put your rook in the corner, immediately, all the spots in here and all the spots in here become invalid for the third rook,

because otherwise, the rooks can hurt each other. So again, you'll see that if you take scissors and cut off the blue part, you will have reduced the dimension of the chessboard again.

And you can see pretty quickly that what you're left with is a 6-by-6 chessboard. So for the third rook, you get a 6-by-6 chessboard, which has 36 free spots. And I'm not going to insult your intelligence. You guys can see the pattern-- 64, 49, 36. These are just perfect squares decreasing.

So you know that the fourth rook will have 25 spots. I'm going to come over here because I'm out of room. The fifth rook will have 16 spots. The sixth rook will have nine spots. The seventh rook will have four spots.

And the eighth rook will just have one spot. And now, here we're going to invoke the counting principle again. Remember the thing that I just defined to you by talking about sandwiches. And we'll see that to get the total number of safe arrangements, we can just multiply these numbers together. So I'm going to go ahead and put that up here.

You get 64 times 49 times 36 times 25 times 16 times 9 times 4. And in fact, this is our answer. So we're all done. So I really like this problem, because we don't normally ask you to think about different spatial arrangements. So it's a nice exercise, because it lets you practice your counting skills in a new and creative way. And in particular, the thing that we've been using for a while now is the discrete uniform law.

But now, I also introduced the counting principle. And we used the counting principle twice-- once to compute the numerator and once to compute the denominator. Counting can take a long time for you to absorb it. So if you still don't totally buy the counting principle, that's OK. I just recommend you do some more examples and try to convince yourself that it's really counting the right number of things.

So counting principle is the second takeaway. And then the other thing that is just worth mentioning is, you guys should get really comfortable with these factorials, because they will just show up again and again. So that's the end of the problem, and I'll see you next time.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Problem Set 3
Due September 29, 2010

1. The hats of n persons are thrown into a box. The persons then pick up their hats at random (i.e., so that every assignment of the hats to the persons is equally likely). What is the probability that
 - (a) every person gets his or her hat back?
 - (b) the first m persons who picked hats get their own hats back?
 - (c) everyone among the first m persons to pick up the hats gets back a hat belonging to one of the last m persons to pick up the hats?

Now assume, in addition, that every hat thrown into the box has probability p of getting dirty (independently of what happens to the other hats or who has dropped or picked it up). What is the probability that

- (d) the first m persons will pick up clean hats?
 - (e) exactly m persons will pick up clean hats?
2. Alice plays with Bob the following game. First Alice randomly chooses 4 cards out of a 52-card deck, memorizes them, and places them back into the deck. Then Bob randomly chooses 8 cards out of the same deck. Alice wins if Bob's cards include all cards selected by her. What is the probability of this happening?
3. (a) Let X be a random variable that takes nonnegative integer values. Show that

$$\mathbf{E}[X] = \sum_{k=1}^{\infty} \mathbf{P}(X \geq k).$$

Hint: Express the right-hand side of the above formula as a double summation then interchange the order of the summations.

- (b) Use the formula in the previous part to find the expectation of a random variable Y whose PMF is defined as follows:

$$p_Y(y) = \frac{1}{b-a+1}, \quad y = a, a+1, \dots, b$$

where a and b are nonnegative integers with $b > a$. Note that for $y = a, a+1, \dots, b$, $p_Y(y)$ does not depend explicitly on y since it is a uniform PMF.

4. Two fair three-sided dice are rolled simultaneously. Let X be the difference of the two rolls.
 - (a) Calculate the PMF, the expected value, and the variance of X .
 - (b) Calculate and plot the PMF of X^2 .
5. Let $n \geq 2$ be an integer. Show that

$$\sum_{k=2}^n k(k-1) \binom{n}{k} = n(n-1)2^{n-2}.$$

Hint: As one way of solving the problem, following from Example 1.31 in the text, think of a committee that includes a chair and a vice-chair.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Fall 2010)

G1[†]. A candy factory has an endless supply of red, orange, yellow, green, blue, black, white, and violet jelly beans. The factory packages the jelly beans into jars in such a way that each jar has 200 beans, equal number of red and orange beans, equal number of yellow and green beans, one more black bean than the number blue beans, and three more violet beans than the number of white beans. One possible color distribution, for example, is a jar of 50 yellow, 50 green, one black, 48 white, and 51 violet jelly beans. As a marketing gimmick, the factory guarantees that no two jars have the same color distribution. What is the maximum number of jars the factory can produce?

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Problem Set 3 Solutions
Due September 29, 2010

1. The hats of n persons are thrown into a box. The persons then pick up their hats at random (i.e., so that every assignment of the hats to the persons is equally likely). What is the probability that

- (a) every person gets his or her hat back?

Answer: $\frac{1}{n!}$.

Solution: consider the sample space of all possible hat assignments. It has $n!$ elements (n hat selections for the first person, after that $n - 1$ for the second, etc.), with every single-element event equally likely (hence having probability $1/n!$). The question is to calculate the probability of a single-element event, so the answer is $1/n!$

- (b) the first m persons who picked hats get their own hats back?

Answer: $\frac{(n-m)!}{n!}$.

Solution: consider the same sample space and probability as in the solution of (a). The probability of an event with $(n - m)!$ elements (this is how many ways there are to distribute the remaining $n - m$ hats after the first m are assigned to their owners) is $(n - m)!/n!$

- (c) everyone among the first m persons to pick up the hats gets back a hat belonging to one of the last m persons to pick up the hats?

Answer: $\frac{m!(n-m)!}{n!} = \frac{1}{\binom{n}{m}} = \frac{1}{\binom{n}{n-m}}$.

Solution: there are $m!$ ways to distribute m hats among the first m persons, and $(n - m)!$ ways to distribute the remaining $n - m$ hats. The probability of an event with $m!(n - m)!$ elements is $m!(n - m)!/n!$.

Now assume, in addition, that every hat thrown into the box has probability p of getting dirty (independently of what happens to the other hats or who has dropped or picked it up). What is the probability that

- (d) the first m persons will pick up clean hats?

Answer: $(1 - p)^m$.

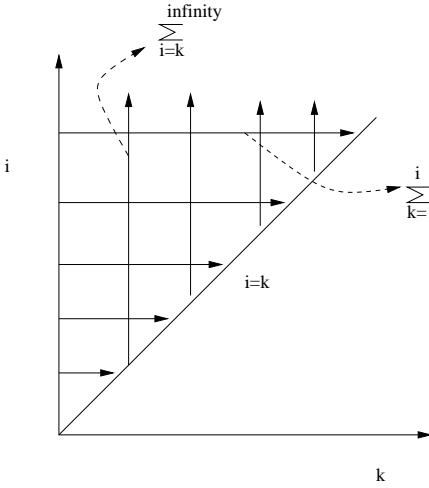
Solution: the probability of a given person picking up a clean hat is $1 - p$. By the independence assumption, the probability of m selected persons picking up clean hats is $(1 - p)^m$.

- (e) exactly m persons will pick up clean hats?

Answer: $(1 - p)^m p^{n-m} \binom{n}{m}$.

Solution: every group G of m persons defines the event “everyone from G picks up a clean hat, everyone not from G picks up a dirty hat”. The events are disjoint. Each has probability $(1 - p)^m p^{n-m}$. Since there are $\binom{n}{m}$ such events, the answer follows.

2. Since 4 cards are fixed, Bob can only choose 4 more cards out of 48 remaining cards, so total number of hands Bob can have such that they include Alice's cards is $\binom{4}{4} \binom{48}{4}$. The total number of ways Bob can choose any 8 cards is $\binom{52}{8}$. So the probability is $\frac{\binom{4}{4} \binom{48}{4}}{\binom{52}{8}}$
3. (a) The picture below illustrates the double sum needed to prove the statement of this problem:



We first note that

$$\mathbf{P}(X \geq k) = \sum_{i=k}^{\infty} p_X(i)$$

and proceed as follows:

$$\sum_{k=1}^{\infty} \mathbf{P}(X \geq k) = \sum_{k=1}^{\infty} \sum_{i=k}^{\infty} p_X(i) = \sum_{i=1}^{\infty} \sum_{k=1}^i p_X(i) = \sum_{i=1}^{\infty} i p_X(i) = \mathbf{E}[X].$$

(b) We first compute

$$\mathbf{P}(Y \geq k) = \begin{cases} 1 & k \leq a \\ \frac{b-k+1}{b-a+1} & a+1 \leq k \leq b \\ 0 & k \geq b+1 \end{cases}$$

So

$$\begin{aligned} \sum_{k=1}^{\infty} \mathbf{P}(Y \geq k) &= \sum_{k=1}^a 1 + \sum_{k=a+1}^b \frac{b-k+1}{b-a+1} \\ &= a + \frac{1}{b-a+1} \sum_{k=1}^{b-a} k \\ &= a + \frac{1}{b-a+1} \frac{(b-a)(b-a+1)}{2} \\ &= a + \frac{b-a}{2} \\ &= \frac{b+a}{2} \end{aligned}$$

Therefore $\mathbf{E}[Y] = \frac{b+a}{2}$.

4. (a) For each value of X , we count the number of outcomes which have a difference that equals that value:

$$p_X(x) = \begin{cases} 1/9 & x = -2, 2 \\ 2/9 & x = -1, 1 \\ 3/9 & x = 0 \\ 0 & \text{otherwise.} \end{cases}$$

$$\mathbf{E}[X] = \sum_{x=-2}^2 x p_X(x) = -2 \frac{1}{9} + -1 \frac{2}{9} + 0 \frac{3}{9} + 1 \frac{2}{9} + 2 \frac{1}{9} = \boxed{0}.$$

We can also see that $\mathbf{E}[X] = 0$ because the PMF is symmetric around 0.

To find the variance of X , we first compute

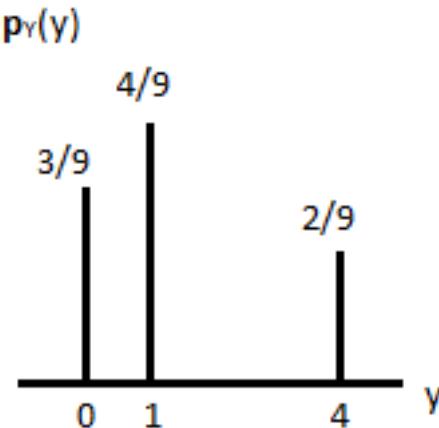
$$\mathbf{E}[X^2] = \sum_{x=-2}^2 x^2 p_X(x) = 4 \frac{1}{9} + 1 \frac{2}{9} + 0 \frac{3}{9} + 1 \frac{2}{9} + 4 \frac{1}{9} = \boxed{\frac{4}{3}}.$$

and

$$\text{var}(X) = \mathbf{E}[X^2] - (\mathbf{E}[X])^2 = \boxed{\frac{4}{3}}.$$

- (b) Let $Z = X^2$. By matching the possible values of X and their probabilities to the possible values of Z , we obtain

$$p_Z(z) = \begin{cases} 2/9 & z = 4 \\ 4/9 & z = 1 \\ 3/9 & z = 0 \\ 0 & \text{otherwise.} \end{cases}$$



5. Consider k out of n persons forming a club, with one being designated as the leader and another as the treasurer. We can first choose the leader (n choices), then the treasurer ($n - 1$ choices), and then a subset of the remaining $n - 2$ persons. Thus, there are $n(n - 1)2^{n-2}$ possible clubs.

Alternatively, for any given k , there are $\binom{n}{k}$ choices for the members of the club. There are $k(k - 1)$ choices for the leader and treasurer, so that there are $k(k - 1)\binom{n}{k}$ k -member clubs. Summing over all k , we see that there is a total of $\sum_{k=2}^n k(k - 1)\binom{n}{k}$ possible clubs.

- G1[†]. A candy factory has an endless supply of red, orange, yellow, green, blue, black, white, and violet jelly beans. The factory packages the jelly beans into jars in such a way that each jar has 200 beans, equal number of red and orange beans, equal number of yellow and green beans, one more black bean than the number blue beans, and three more violet beans than the number of white

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
 (Fall 2010)

beans. One possible color distribution, for example, is a jar of 50 yellow, 50 green, one black, 48 white, and 51 violet jelly beans. As a marketing gimmick, the factory guarantees that no two jars have the same color distribution. What is the maximum number of jars the factory can produce?

Answer: $\binom{101}{3} = 166650$.

Solution: Let $N_1, N_2, N_3, N_4, N_5, N_6, N_7, N_8$ denote, respectively, the numbers of red, orange, yellow, green, blue, black, white, and violet jelly beans in a jar. There is a one-to-one correspondence

$$x = (x_1, x_2, x_3, x_4) \mapsto N = (x_1, x_1, x_2, x_2, x_3, x_3 + 1, x_4, x_4 + 3)$$

between the non-negative integer solutions $x = (x_1, x_2, x_3, x_4)$ of the equation

$$x_1 + x_2 + x_3 + x_4 = 98,$$

and the sequences $N = (N_1, N_2, N_3, N_4, N_5, N_6, N_7, N_8)$ of non-negative integers N_i satisfying the conditions

$$N_2 = N_1, \quad N_4 = N_3, \quad N_6 = N_5 + 1, \quad N_8 = N_7 + 3, \quad \sum_{i=1}^8 N_i = 200$$

(i.e. possible color arrangements). The number of possible solutions x is $\binom{101}{3}$ according to the solution of the more general problem given below:

Given a non-negative integer n and a positive integer k , consider the equation

$$x_1 + x_2 + \dots + x_k = n,$$

to be solved with respect to non-negative integer variables x_1, x_2, \dots, x_k . Find the total number of solutions (solutions $x_1 = 1, x_2 = 0$ and $x_1 = 0, x_2 = 1$ to the equation $x_1 + x_2 = 1$ are considered as different).

Answer: $\binom{n+k-1}{k-1} = \binom{n+k-1}{n}$.

Solution: there is a one-to-one correspondence between non-negative integer solutions of equation $x_1 + \dots + x_k = n$ and sequences of $n+k-1$ symbols (n “o” and $k-1$ “|”), where a solution $x = (x_1, \dots, x_k)$ maps to the sequence in which the i -th “|” (where $i \in \{1, 2, \dots, k-1\}$) is in the $x_1 + \dots + x_i + i$ th place: in this bijection, the numbers of “o” between the consecutive “|” correspond to the values of x_i . Hence the total number of solutions equals the number of ways of selecting $k-1$ places for the “|” symbols in a sequence of length $n+k-1$.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

LECTURE 4

- **Readings:** Section 1.6

Lecture outline

- Principles of counting
- Many examples
 - permutations
 - k -permutations
 - combinations
 - partitions
- Binomial probabilities

Discrete uniform law

- Let all sample points be equally likely

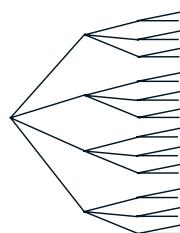
- Then,

$$P(A) = \frac{\text{number of elements of } A}{\text{total number of sample points}} = \frac{|A|}{|\Omega|}$$

- Just count...

Basic counting principle

- r stages
- n_i choices at stage i



- Number of choices is: $n_1 n_2 \cdots n_r$
- Number of license plates with 3 letters and 4 digits =
- ... if repetition is prohibited =
- **Permutations:** Number of ways of ordering n elements is:
- Number of subsets of $\{1, \dots, n\}$ =

Example

- Probability that six rolls of a six-sided die all give different numbers?
 - Number of outcomes that make the event happen:
 - Number of elements in the sample space:
 - Answer:

Combinations

- $\binom{n}{k}$: number of k -element subsets of a given n -element set
- Two ways of constructing an ordered sequence of k **distinct** items:
 - Choose the k items one at a time:
 $n(n-1)\cdots(n-k+1) = \frac{n!}{(n-k)!}$ choices
 - Choose k items, then order them ($k!$ possible orders)
- Hence:

$$\binom{n}{k} \cdot k! = \frac{n!}{(n-k)!}$$

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

$$\sum_{k=0}^n \binom{n}{k} =$$

Binomial probabilities

- n independent coin tosses
 - $P(H) = p$
- $P(HTTHHH) =$
- $P(\text{sequence}) = p^{\# \text{ heads}}(1-p)^{\# \text{ tails}}$

$$\begin{aligned} P(k \text{ heads}) &= \sum_{k-\text{head seq.}} P(\text{seq.}) \\ &= (\# \text{ of } k\text{-head seqs.}) \cdot p^k (1-p)^{n-k} \\ &= \binom{n}{k} p^k (1-p)^{n-k} \end{aligned}$$

Coin tossing problem

- event B : 3 out of 10 tosses were “heads”
 - Given that B occurred, what is the (conditional) probability that the first 2 tosses were heads?
- All outcomes in set B are equally likely: probability $p^3(1-p)^7$
- Conditional probability law is uniform
- Number of outcomes in B :
- Out of the outcomes in B , how many start with HH?

Partitions

- 52-card deck, dealt to 4 players
- Find $P(\text{each gets an ace})$
- Outcome: a partition of the 52 cards
 - number of outcomes:

$$\frac{52!}{13! 13! 13! 13!}$$
- Count number of ways of distributing the four aces: $4 \cdot 3 \cdot 2$
- Count number of ways of dealing the remaining 48 cards

$$\frac{48!}{12! 12! 12! 12!}$$

- Answer:

$$\frac{4 \cdot 3 \cdot 2}{12! 12! 12! 12!} \frac{48!}{52!} \frac{1}{13! 13! 13! 13!}$$

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Recitation 4
September 21, 2010

1. Problem 1.50, page 67 in the text.

The birthday problem. Consider n people who are attending a party. We assume that every person has an equal probability of being born on any day during the year, independently of everyone else, and ignore the additional complication presented by leap years (i.e., nobody is born on February 29). What is the probability that each person has a distinct birthday?

2. Imagine that 8 rooks are randomly placed on a chessboard. Find the probability that all the rooks will be safe from one another, i.e. that there is no row or column with more than one rook.
3. Problem 1.61, page 69 in the text.

Hypergeometric probabilities. An urn contains n balls, out of which exactly m are red. We select k of the balls at random, without replacement (i.e., selected balls are not put back into the urn before the next selection). What is the probability that i of the selected balls are red?

4. **Multinomial coefficient.** Derive the multinomial coefficient (the number of partitions of n distinct items into groups of n_1, \dots, n_r) using a different argument than the one in class. Consider n items which can be placed into n slots and divide the group of n slots into segments of length n_1, \dots, n_r slots. Derive the multinomial coefficient by showing how many different ways can the n items be arranged into the r segments.
5. **Multinomial probabilities.** At each draw, there is a probability p_i ($i = 1, \dots, r$) of getting a ball of color i . Draw n objects. What is the probability of obtaining exactly n_i of each color i ?

Recitation 4: Extra Handout
September 21, 2010

- As part of the solution to problem 1, plotted below are the probabilities of each person having a distinct birthday versus n the number of people present.

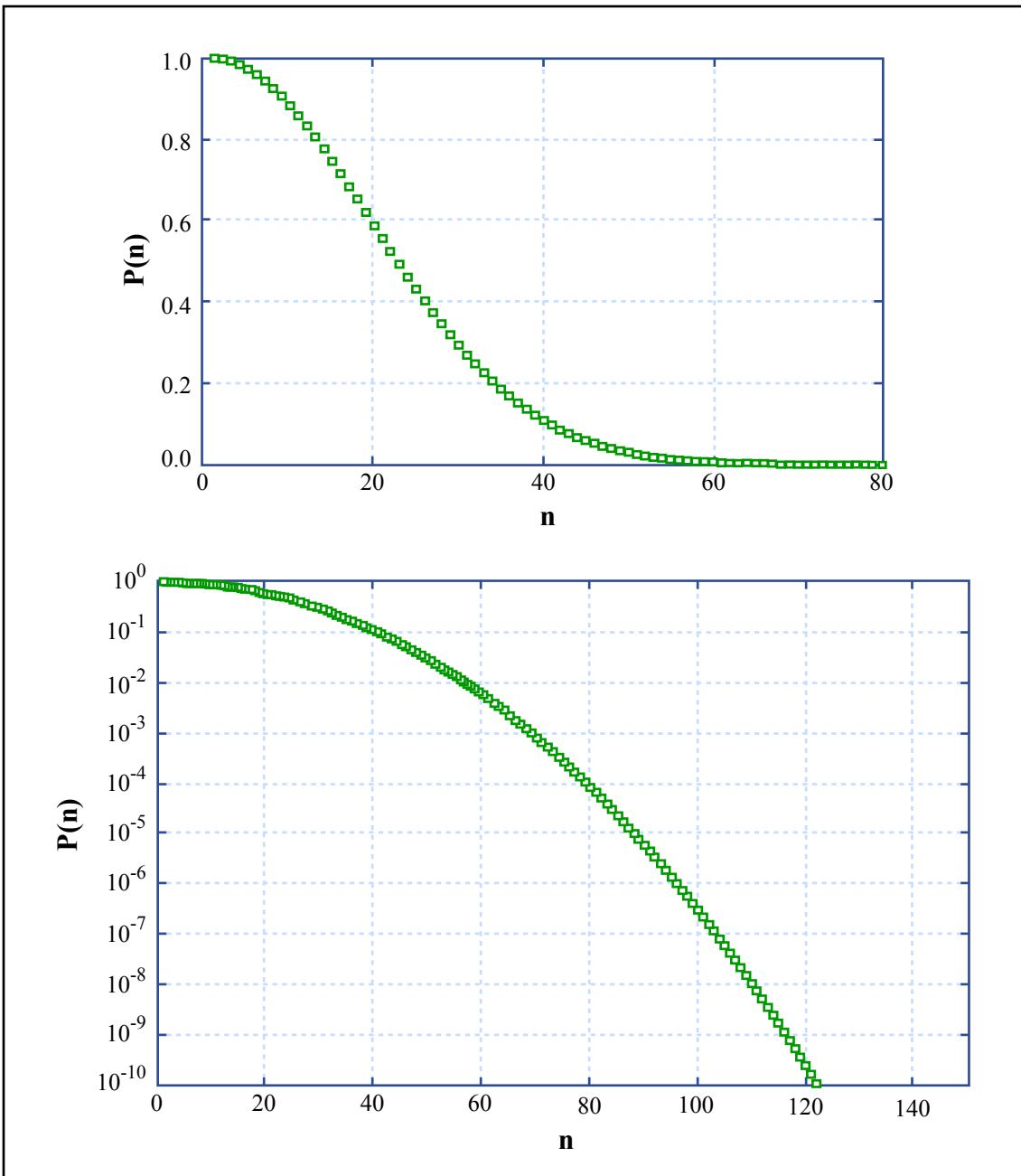


Image by MIT OpenCourseWare.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Recitation 4 Solutions
September 21, 2010

1. The sample space consists of all possible choices for the birthday of each person. Since there are n persons, and each has 365 choices for their birthday, the sample space has 365^n elements. Let us now consider those choices of birthdays for which no two persons have the same birthday. Assuming that $n \leq 365$, there are 365 choices for the first person, 364 for the second, etc., for a total of $365 \cdot 364 \cdots (365 - n + 1)$. Thus,

$$P(\text{no two birthdays coincide}) = \frac{365 \cdot 364 \cdots (365 - n + 1)}{365^n}.$$

It is interesting to note that for n as small as 23, the probability that there are two persons with the same birthday is larger than 1/2.

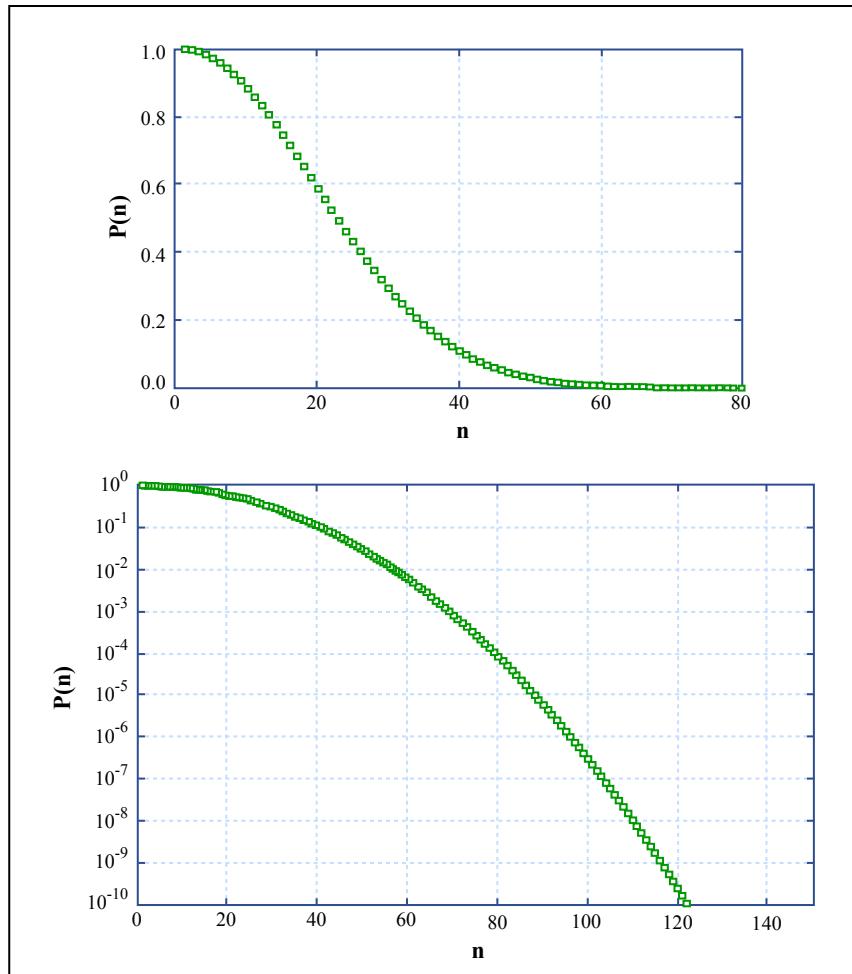


Image by MIT OpenCourseWare.

2. As we have done before, we will count the number of favorable positions in which we can safely place 8 rooks, and then divide this by the total number of positions for 8 rooks on a 8×8 chessboard. First we count the number of favorable positions for the rooks. We will place the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Fall 2010)

rooks one by one. For the first rook, there are no constraints, so we have 64 choices. Placing this rook, however, eliminates one row and one column. Thus for our second rook, we can imagine that the illegal column and row have been removed, thus leaving us with a 7×7 chessboard, and thus with 49 choices. Similarly, for the third rook we have 36 choices, for the fourth 25, etc... There are $64 \cdot 63 \cdots 57$ total ways we can place 8 rooks without any restrictions, and therefore the probability we are after is:

$$\frac{64 \cdot 49 \cdot 36 \cdot 25 \cdot 16 \cdot 9 \cdot 4}{\frac{64!}{56!}}.$$

3. See textbook, Problem 1.61, page 69.
4. The group of n slots is divided into segments of length n_1, \dots, n_r slots. The n items can be arranged in $n!$ ways, where each arrangement corresponds to a partition into the r segments. But all arrangements within a single segment lead to the same partition, where there are $n_i!$ ways to arrange the items within i th segment. Thus, for each segment we must divide by the number of ways to arrange the items within that segment. The solution is then:

$$\frac{\text{Ways to arrange } n \text{ items}}{(\text{Ways to arrange items in segment 1}) \cdots (\text{Ways to arrange items in segment } r)} = \frac{n!}{n_1! \cdots n_r!}$$

5. The probability of drawing a particular sequence of balls containing exactly n_i of color i balls is $p_1^{n_1} \cdots p_r^{n_r}$. The number of possible sequences containing n_i of color i balls is the number of ways to form a partition of n distinct slots into subsets of cardinality n_1, \dots, n_r which is $\binom{n}{n_1, \dots, n_r}$. Therefore, the probability of obtaining exactly n_i balls of color i is:

$$\binom{n}{n_1, \dots, n_r} p_1^{n_1} \cdots p_r^{n_r}$$

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Tutorial 2
September 23/24, 2010

1. A player is randomly dealt 13 cards from a standard 52-card deck.
 - (a) What is the probability the 13th card dealt is a king?
 - (b) What is the probability the 13th card dealt is the first king dealt?

2. Consider a random variable X such that

$$p_X(x) = \frac{x^2}{a} \text{ for } x \in \{-3, -2, -1, 1, 2, 3\}, \quad \mathbf{P}(X = x) = 0 \text{ for } x \notin \{-3, -2, -1, 1, 2, 3\},$$

where $a > 0$ is a real parameter.

- (a) Find a .
 - (b) What is the PMF of the random variable $Z = X^2$?
3. 90 students, including Joe and Jane, are to be split into three classes of equal size, and this is to be done at random. What is the probability that Joe and Jane end up in the same class?
 4. Draw the top 7 cards from a well-shuffled standard 52-card deck. Find the probability that the 7 cards include exactly 3 aces.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Tutorial 2 Solutions
September 23/24, 2010

1. A player is randomly dealt 13 cards from a standard 52-card deck.

- (a) What is the probability the 13th card dealt is a king?

Answer: $\frac{4}{52}$.

Solution: Since we are not told anything about the first 12 cards that are dealt, the probability that the 13th card dealt is a King, is the same as the probability that the first card dealt, or in fact any particular card dealt is a King, and this equals: $\frac{4}{52}$.

- (b) What is the probability the 13th card dealt is the first king dealt?

Answer: $\frac{1}{13} \cdot 4 \binom{48}{12} / \binom{52}{13}$.

Solution: The probability that the 13th card is the first king to be dealt is the probability that out of the first 13 cards to be dealt, exactly one was a king, and that the king was dealt last. Now, given that exactly one king was dealt in the first 13 cards, the probability that the king was dealt last is just $1/13$, since each “position” is equally likely. Thus, it remains to calculate the probability that there was exactly one king in the first 13 cards dealt. To calculate this probability we count the “favorable” outcomes and divide by the total number of possible outcomes. We first count the favorable outcomes, namely those with exactly one king in the first 13 cards dealt. We can choose a particular king in 4 ways, and we can choose the other 12 cards in $\binom{48}{12}$ ways, therefore there are $4 \cdot \binom{48}{12}$ favorable outcomes. There are $\binom{52}{13}$ total outcomes, so the desired probability is

$$\frac{1}{13} \cdot \frac{4 \binom{48}{12}}{\binom{52}{13}}.$$

For an alternative solution, we argue as in Example 1.10. The probability that the first card is not a king is $48/52$. Given that, the probability that the second is not a king is $47/51$. We continue similarly until the 12th card. The probability that the 12th card is not a king, given that none of the preceding 11 was a king, is $37/41$. (There are $52 - 11 = 41$ cards left, and $48 - 11 = 37$ of them are not kings.) Finally, the conditional probability that the 13th card is a king is $4/40$. The desired probability is

$$\frac{48 \cdot 47 \cdots 37 \cdot 4}{52 \cdot 51 \cdots 41 \cdot 40}.$$

2. Consider a random variable X such that

$$p_X(x) = \frac{x^2}{a} \text{ for } x \in \{-3, -2, -1, 1, 2, 3\}, \quad \mathbf{P}(X = x) = 0 \text{ for } x \notin \{-3, -2, -1, 1, 2, 3\},$$

where $a > 0$ is a real parameter.

- (a) Find a .

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
 (Fall 2010)

Solution. The sum of the values of the PMF of a random variable over all values that it takes with positive probability must be equal to 1. Hence, we have

$$\begin{aligned} 1 &= \sum_{x=-3}^3 p_X(x) \\ &= \frac{9}{a} + \frac{4}{a} + \frac{1}{a} + \frac{1}{a} + \frac{4}{a} + \frac{9}{a} \\ &= \frac{28}{a}, \end{aligned}$$

which implies that $a = 28$.

- (b) What is the PMF of the random variable $Z = X^2$?

Solution. The following table shows the value of Z for a given value of X and the probability of that event.

x	-3	-2	-1	1	2	3
$p_X(x)$	9/28	1/7	1/28	1/28	1/7	9/28
$Z X=x$	9	4	1	1	4	9

We see that Z can take only three possible values with non-zero probability, namely 1, 4, and 9. In addition, for each value, there correspond two values of X . So we have, for example, $p_Z(9) = \mathbf{P}(Z = 9) = \mathbf{P}(X = -3) + \mathbf{P}(X = 3) = p_X(-3) + p_X(3)$. Hence the PMF of Z is given by

$$p_Z(z) = \begin{cases} 1/14 & \text{if } z = 1, \\ 2/7 & \text{if } z = 4, \\ 9/14 & \text{if } z = 9. \end{cases}$$

3. Suppose we label the classes A , B , and C . Now the probability that Joe and Jane will both be in class A is the number of possible combinations for class A that involve both Joe and Jane, divided by the total number of combinations for class A . Therefore the probability we are after is:

$$\frac{\binom{88}{28}}{\binom{90}{30}}.$$

Since there are three classrooms, the probability that Joe and Jane end up in the same classroom is simply three times the answer we found above:

$$3 \cdot \frac{\binom{88}{28}}{\binom{90}{30}}.$$

Another way of looking at the problem is described as follows,

Assume one of them pick first, say Joe. He can pick any one of the 90 available places. Then it's Jane's turn to pick. She has a probability of $\frac{29}{89}$ of picking in the same class as Joe. Therefore, the overall probability is $\frac{29}{89}$, which is the same as $3 \cdot \frac{\binom{88}{28}}{\binom{90}{30}}$.

4. Let A = event the 7 cards include exactly 3 aces.

$$P(A) = \frac{(\# \text{ ways to choose 3 aces}) \cdot (\# \text{ ways to choose other 4 cards})}{\# \text{ ways to choose 7 cards}} = \frac{\binom{4}{3} \binom{48}{4}}{\binom{52}{7}}.$$

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013 Transcript – Lecture 5

The following content is provided under a Creative Commons license. Your support will help MIT OpenCourseWare continue to offer high quality educational resources for free. To make a donation, or view additional materials from hundreds of MIT courses, visit MIT OpenCourseWare at ocw.mit.edu,

OK. So let us start. All right. So today we're starting a new unit in this class. We have covered, so far, the basics of probability theory-- the main concepts and tools, as far as just probabilities are concerned. But if that was all that there is in this subject, the subject would not be rich enough. What makes probability theory a lot more interesting and richer is that we can also talk about random variables, which are ways of assigning numerical results to the outcomes of an experiment.

So we're going to define what random variables are, and then we're going to describe them using so-called probability mass functions. Basically some numerical values are more likely to occur than other numerical values, and we capture this by assigning probabilities to them the usual way. And we represent these in a compact way using the so-called probability mass functions.

We're going to see a couple of examples of random variables, some of which we have already seen but with different terminology. And so far, it's going to be just a couple of definitions and calculations of the type that you already know how to do. But then we're going to introduce the one new, big concept of the day.

So up to here it's going to be mostly an exercise in notation and definitions. But then we got our big concept which is the concept of the expected value of a random variable, which is some kind of average value of the random variable. And then we're going to also talk, very briefly, about close distance of the expectation, which is the concept of the variance of a random variable.

OK. So what is a random variable? It's an assignment of a numerical value to every possible outcome of the experiment. So here's the picture. The sample space is this class, and we've got lots of students in here. This is our sample space, Ω . I'm interested in the height of a random student. So I'm going to use a real line where I record height, and let's say this is height in inches.

And the experiment happens, I pick a random student. And I go and measure the height of that random student, and that gives me a specific number. So what's a good number in inches? Let's say 60. OK. Or I pick another student, and that student has a height of 71 inches, and so on.

So this is the experiment. These are the outcomes. These are the numerical values of the random variable that we call height. OK. So mathematically, what are we dealing with here? We're basically dealing with a function from the sample space into the real numbers. That function takes as argument, an outcome of the experiment, that is a typical student, and produces the value of that function, which is the height of that particular student.

So we think of an abstract object that we denote by a capital H, which is the random variable called height. And that random variable is essentially this particular function that we talked about here. OK. So there's a distinction that we're making here-- H is height in the abstract. It's the function. These numbers here are particular numerical values that this function takes when you choose one particular outcome of the experiment.

Now, when you have a single probability experiment, you can have multiple random variables. So perhaps, instead of just height, I'm also interested in the weight of a typical student. And so when the experiment happens, I pick that random student-- this is the height of the student. But that student would also have a weight, and I could record it here. And similarly, every student is going to have their own particular weight.

So the weight function is a different function from the sample space to the real numbers, and it's a different random variable. So the point I'm making here is that a single probabilistic experiment may involve several interesting random variables. I may be interested in the height of a random student or the weight of the random student. These are different random variables that could be of interest.

I can also do other things. Suppose I define an object such as $H\bar{}$, which is 2.58. What does that correspond to? Well, this is the height in centimeters. Now, $H\bar{}$ is a function of H itself, but if you were to draw the picture, the picture would go this way. 60 gets mapped to 150, 71 gets mapped to, oh, that's too hard for me. OK, gets mapped to something, and so on.

So $H\bar{}$ is also a random variable. Why? Once I pick a particular student, that particular outcome determines completely the numerical value of $H\bar{}$, which is the height of that student but measured in centimeters. What we have here is actually a random variable, which is defined as a function of another random variable. And the point that this example is trying to make is that functions of random variables are also random variables.

The experiment happens, the experiment determines a numerical value for this object. And once you have the numerical value for this object, that determines also the numerical value for that object. So given an outcome, the numerical value of this particular object is determined. So $H\bar{}$ is itself a function from the sample space, from outcomes to numerical values. And that makes it a random variable according to the formal definition that we have here.

So the formal definition is that the random variable is not random, it's not a variable, it's just a function from the sample space to the real numbers. That's the abstract, right way of thinking about them. Now, random variables can be of different types. They can be discrete or continuous.

Suppose that I measure the heights in inches, but I round to the nearest inch. Then the numerical values I'm going to get here would be just integers. So that would make it an integer valued random variable. And this is a discrete random variable.

Or maybe I have a scale for measuring height which is infinitely precise and records your height to an infinite number of digits of precision. In that case, your height would be just a general real

number. So we would have a random variable that takes values in the entire set of real numbers. Well, I guess not really negative numbers, but the set of non-negative numbers. And that would be a continuous random variable. It takes values in a continuous set.

So we will be talking about both discrete and continuous random variables. The first thing we will do will be to devote a few lectures on discrete random variables, because discrete is always easier. And then we're going to repeat everything in the continuous setting. So discrete is easier, and it's the right place to understand all the concepts, even those who may appear to be elementary. And then you will be set to understand what's going on when we go to the continuous case.

So in the continuous case, you get all the complications of calculus and some extra math that comes in there. So it's important to have been down all the concepts very well in the easy, discrete case so that you don't have conceptual hurdles when you move on to the continuous case.

Now, one important remark that may seem trivial but it's actually very important so that you don't get tangled up between different types of concepts-- there's a fundamental distinction between the random variable itself, and the numerical values that it takes. Abstractly speaking, or mathematically speaking, a random variable, x , or H in this example, is a function.

OK. Maybe if you like programming the words "procedure" or "sub-routine" might be better. So what's the sub-routine height? Given a student, I take that student, force them on the scale and measure them. That's the sub-routine that measures heights. It's really a function that takes students as input and produces numbers as output.

The sub-routine we denoted by capital H . That's the random variable. But once you plug in a particular student into that sub-routine, you end up getting a particular number. This is the numerical output of that sub-routine or the numerical value of that function. And that numerical value is an element of the real numbers.

So the numerical value is a real number, where this capital X is a function from Ω to the real numbers. So they are very different types of objects. And the way that we keep track of what we're talking about at any given time is by using capital letters for random variables and lower case letters for numbers.

OK. So now once we have a random variable at hand, that random variable takes on different numerical values. And we want to describe to say something about the relative likelihoods of the different numerical values that the random variable can take.

So here's our sample space, and here's the real line. And there's a bunch of outcomes that gave rise to one particular numerical value. There's another numerical value that arises if we have this outcome. There's another numerical value that arises if we have this outcome. So our sample space is here. The real numbers are here. And what we want to do is to ask the question, how likely is that particular numerical value to occur?

So what we're essentially asking is, how likely is it that we obtain an outcome that leads to that particular numerical value? We calculate that overall probability of that numerical value and we represent that probability using a bar so that we end up generating a bar graph. So that could be a possible bar graph associated with this picture. The size of this bar is the total probability that our random variable took on this numerical value, which is just the sum of the probabilities of the different outcomes that led to that numerical value.

So the thing that we're plotting here, the bar graph-- we give a name to it. It's a function, which we denote by lowercase b , capital X . The capital X indicates which random variable we're talking about. And it's a function of little x , which is the range of values that our random variable is taking.

So in mathematical notation, the value of the PMF at some particular number, little x , is the probability that our random variable takes on the numerical value, little x . And if you want to be precise about what this means, it's the overall probability of all outcomes for which the random variable ends up taking that value, little x .

So this is the overall probability of all omegas that lead to that particular numerical value, x , of interest. So what do we know about PMFs? Since there are probabilities, all these entries in the bar graph have to be non-negative. Also, if you exhaust all the possible values of little x 's, you will have exhausted all the possible outcomes here. Because every outcome leads to some particular x .

So the sum of these probabilities should be equal to one. This is the second relation here. So this relation tell us that some little x is going to happen. They happen with different probabilities, but when you consider all the possible little x 's together, one of those little x 's is going to be realized. Probabilities need to add to one.

OK. So let's get our first example of a non-trivial bar graph. Consider the experiment where I start with a coin and I start flipping it over and over. And I do this until I obtain heads for the first time. So what are possible outcomes of this experiment?

One possible outcome is that I obtain heads at the first toss, and then I stop. In this case, my random variable takes the value 1. Or it's possible that I obtain tails and then heads. How many tosses did it take until heads appeared? This would be x equals to 2. Or more generally, I might obtain tails for k minus 1 times, and then obtain heads at the k -th time, in which case, our random variable takes the value, little k .

So that's the experiment. So capital X is a well defined random variable. It's the number of tosses it takes until I see heads for the first time. These are the possible outcomes. These are elements of our sample space. And these are the values of X depending on the outcome. Clearly X is a function of the outcome. You tell me the outcome, I'm going to tell you what X is.

So what we want to do now is to calculate the PMF of X . So P_x of k is, by definition, the probability that our random variable takes the value k . For the random variable to take the value

of k , the first head appears at toss number k . The only way that this event can happen is if we obtain this sequence of events.

T 's the first k minus 1 times, tails, and heads at the k -th flip. So this event, that the random variable is equal to k , is the same as this event, k minus 1 tails followed by 1 head. What's the probability of that event?

We're assuming that the coin tosses are independent. So to find the probability of this event, we need to multiply the probability of tails, times the probability of tails, times the probability of tails. We multiply k minus one times, times the probability of heads, which puts an extra p at the end. And this is the formula for the so-called geometric PMF.

And why do we call it geometric? Because if you go and plot the bar graph of this random variable, X , we start at 1 with a certain number, which is p . And then at 2 we get $p(1-p)$. At 3 we're going to get something smaller, it's p times $(1-p)$ -squared. And the bars keep going down at the rate of geometric progression. Each bar is smaller than the previous bar, because each time we get an extra factor of $1-p$ involved.

So the shape of this PMF is the graph of a geometric sequence. For that reason, we say that it's the geometric PMF, and we call X also a geometric random variable. So the number of coin tosses until the first head is a geometric random variable.

So this was an example of how to compute the PMF of a random variable. This was an easy example, because this event could be realized in one and only one way. So to find the probability of this, we just needed to find the probability of this particular outcome. More generally, there's going to be many outcomes that can lead to the same numerical value. And we need to keep track of all of them.

For example, in this picture, if I want to find this value of the PMF, I need to add up the probabilities of all the outcomes that leads to that value. So the general procedure is exactly what this picture suggests. To find this probability, you go and identify which outcomes lead to this numerical value, and add their probabilities.

So let's do a simple example. I take a tetrahedral die. I toss it twice. And there's lots of random variables that you can associate with the same experiment. So the outcome of the first throw, we can call it F . That's a random variable because it's determined once you tell me what happens in the experiment.

The outcome of the second throw is another random variable. The minimum of the two throws is also a random variable. Once I do the experiment, this random variable takes on a specific numerical value. So suppose I do the experiment and I get a 2 and a 3. So this random variable is going to take the numerical value of 2. This is going to take the numerical value of 3. This is going to take the numerical value of 2.

And now suppose that I want to calculate the PMF of this random variable. What I will need to do is to calculate $P_x(0)$, $P_x(1)$, $P_x(2)$, $P_x(3)$, and so on. Let's not do the entire calculation then, let's just calculate one of the entries of the PMF.

So $P_x(2)$ -- that's the probability that the minimum of the two throws gives us a 2. And this can happen in many ways. There are five ways that it can happen. Those are all of the outcomes for which the smallest of the two is equal to 2. That's five outcomes assuming that the tetrahedral die is fair and the tosses are independent. Each one of these outcomes has probability of $1/16$. There's five of them, so we get an answer, $5/16$.

Conceptually, this is just the procedure that you use to calculate PMFs the way that you construct this particular bar graph. You consider all the possible values of your random variable, and for each one of those random variables you find the probability that the random variable takes on that value by adding the probabilities of all the possible outcomes that leads to that particular numerical value.

So let's do another, more interesting one. So let's revisit the coin tossing problem from last time. Let us fix a number n , and we decide to flip a coin n consecutive times. Each time the coin tosses are independent. And each one of the tosses will have a probability, p , of obtaining heads.

Let's consider the random variable, which is the total number of heads that have been obtained. Well, that's something that we dealt with last time. We know the probabilities for different numbers of heads, but we're just going to do the same now using today's notation.

So let's, for concreteness, n equal to 4. P_x is the PMF of that random variable, X . $P_x(2)$ is meant to be, by definition, it's the probability that a random variable takes the value of 2. So this is the probability that we have, exactly two heads in our four tosses.

The event of exactly two heads can happen in multiple ways. And here I've written down the different ways that it can happen. It turns out that there's exactly six ways that it can happen. And each one of these ways, luckily enough, has the same probability-- p -squared times $(1-p)$ -squared. So that gives us the value for the PMF evaluated at 2.

So here we just counted explicitly that we have six possible ways that this can happen, and this gave rise to this factor of 6. But this factor of 6 turns out to be the same as this 4 choose 2. If you remember definition from last time, 4 choose 2 is 4 factorial divided by 2 factorial, divided by 2 factorial, which is indeed equal to 6. And this is the more general formula that you would be using.

In general, if you have n tosses and you're interested in the probability of obtaining k heads, the probability of that event is given by this formula. So that's the formula that we derived last time. Except that last time we didn't use this notation. We just said the probability of k heads is equal to this. Today we introduce the extra notation.

And also having that notation, we may be tempted to also plot a bar graph for the P_x . In this case, for the coin tossing problem. And if you plot that bar graph as a function of k when n is a

fairly large number, what you will end up obtaining is a bar graph that has a shape of something like this.

So certain values of k are more likely than others, and the more likely values are somewhere in the middle of the range. And extreme values-- too few heads or too many heads, are unlikely. Now, the miraculous thing is that it turns out that this curve gets a pretty definite shape, like a so-called bell curve, when n is big.

This is a very deep and central fact from probability theory that we will get to in a couple of months. For now, it just could be a curious observation. If you go into MATLAB and put this formula in and ask MATLAB to plot it for you, you're going to get an interesting shape of this form. And later on we will have to sort of understand where this is coming from and whether there's a nice, simple formula for the asymptotic form that we get.

All right. So, so far I've said essentially nothing new, just a little bit of notation and this little conceptual thing that you have to think of random variables as functions in the sample space. So now it's time to introduce something new. This is the big concept of the day. In some sense it's an easy concept.

But it's the most central, most important concept that we have to deal with random variables. It's the concept of the expected value of a random variable. So the expected value is meant to be, let's speak loosely, something like an average, where you interpret probabilities as something like frequencies.

So you play a certain game and your rewards are going to be-- use my standard numbers-- your rewards are going to be one dollar with probability $1/6$. It's going to be 2 dollars with probability $1/2$, and four dollars with probability $1/3$. So this is a plot of the PMF of some random variable. If you play that game and you get so many dollars with this probability, and so on, how much do you expect to get on the average if you play the game a zillion times?

Well, you can think as follows-- one sixth of the time I'm going to get one dollar. One half of the time that outcome is going to happen and I'm going to get two dollars. And one third of the time the other outcome happens, and I'm going to get four dollars. And you evaluate that number and it turns out to be 2.5. OK. So that's a reasonable way of calculating the average payoff if you think of these probabilities as the frequencies with which you obtain the different payoffs.

And loosely speaking, it doesn't hurt to think of probabilities as frequencies when you try to make sense of various things. So what did we do here? We took the probabilities of the different outcomes, of the different numerical values, and multiplied them with the corresponding numerical value.

Similarly here, we have a probability and the corresponding numerical value and we added up over all x 's. So that's what we did. It looks like an interesting quantity to deal with. So we're going to give a name to it, and we're going to call it the expected value of a random variable. So this formula just captures the calculation that we did.

How do we interpret the expected value? So the one interpretation is the one that I used in this example. You can think of it as the average that you get over a large number of repetitions of an experiment where you interpret the probabilities as the frequencies with which the different numerical values can happen.

There's another interpretation that's a little more visual and that's kind of insightful, if you remember your freshman physics, this kind of formula gives you the center of gravity of an object of this kind. If you take that picture literally and think of this as a mass of one sixth sitting here, and the mass of one half sitting here, and one third sitting there, and you ask me what's the center of gravity of that structure. This is the formula that gives you the center of gravity.

Now what's the center of gravity? It's the place where if you put your pen right underneath, that diagram will stay in place and will not fall on one side and will not fall on the other side. So in this thing, by picture, since the 4 is a little more to the right and a little heavier, the center of gravity should be somewhere around here. And that's what for the math gave us. It turns out to be two and a half.

Once you have this interpretation about centers of gravity, sometimes you can calculate expectations pretty fast. So here's our new random variable. It's the uniform random variable in which each one of the numerical values is equally likely. Here there's a total of $n + 1$ possible numerical values. So each one of them has probability $1 \text{ over } (n + 1)$.

Let's calculate the expected value of this random variable. We can take the formula literally and consider all possible outcomes, or all possible numerical values, and weigh them by their corresponding probability, and do this calculation and obtain an answer. But I gave you the intuition of centers of gravity. Can you use that intuition to guess the answer?

What's the center of gravity infrastructure of this kind? We have symmetry. So it should be in the middle. And what's the middle? It's the average of the two end points. So without having to do the algebra, you know that's the answer is going to be $n \text{ over } 2$.

So this is a moral that you should keep whenever you have PMF, which is symmetric around a certain point. That certain point is going to be the expected value associated with this particular PMF. OK. So having defined the expected value, what is there that's left for us to do?

Well, we want to investigate how it behaves, what kind of properties does it have, and also how do you calculate expected values of complicated random variables. So the first complication that we're going to start with is the case where we deal with a function of a random variable.

OK. So let me redraw this same picture as before. We have ω . This is our sample space. This is the real line. And we have a random variable that gives rise to various values for X . So the random variable is capital X , and every outcome leads to a particular numerical value x for our random variable X . So capital X is really the function that maps these points into the real line.

And then I consider a function of this random variable, call it capital Y, and it's a function of my previous random variable. And this new random variable Y takes numerical values that are completely determined once I know the numerical value of capital X. And perhaps you get a diagram of this kind.

So X is a random variable. Once you have an outcome, this determines the value of x. Y is also a random variable. Once you have the outcome, that determines the value of y. Y is completely determined once you know X. We have a formula for how to calculate the expected value of X.

Suppose that you're interested in calculating the expected value of Y. How would you go about it? OK. The only thing you have in your hands is the definition, so you could start by just using the definition. And what does this entail? It entails for every particular value of y, collect all the outcomes that leads to that value of y. Find their probability. Do the same here. For that value, collect those outcomes. Find their probability and weight by y.

So this formula does the addition over this line. We consider the different outcomes and add things up. There's an alternative way of doing the same accounting where instead of doing the addition over those numbers, we do the addition up here. We consider the different possible values of x, and we think as follows-- for each possible value of x, that value is going to occur with this probability. And if that value has occurred, this is how much I'm getting, the g of x.

So I'm considering the probability of this outcome. And in that case, y takes this value. Then I'm considering the probabilities of this outcome. And in that case, g of x takes again that value. Then I consider this particular x, it happens with this much probability, and in that case, g of x takes that value, and similarly here.

We end up doing exactly the same arithmetic, it's only a question whether we bundle things together. That is, if we calculate the probability of this, then we're bundling these two cases together. Whereas if we do the addition up here, we do a separate calculation-- this probability times this number, and then this probability times that number.

So it's just a simple rearrangement of the way that we do the calculations, but it does make a big difference in practice if you actually want to calculate expectations. So the second procedure that I mentioned, where you do the addition by running over the x-axis corresponds to this formula. Consider all possibilities for x and when that x happens, how much money are you getting? That gives you the average money that you are getting.

All right. So I kind of hand waved and argued that it's just a different way of accounting, of course one needs to prove this formula. And fortunately it can be proved. You're going to see that in recitation. Most people, once they're a little comfortable with the concepts of probability, actually believe that this is true by definition. In fact it's not true by definition. It's called the law of the unconscious statistician. It's something that you always do, but it's something that does require justification.

All right. So this gives us basically a shortcut for calculating expected values of functions of a random variable without having to find the PMF of that function. We can work with the PMF of the original function. All right. So we're going to use this property over and over.

Before we start using it, one general word of caution-- the average of a function of a random variable, in general, is not the same as the function of the average. So these two operations of taking averages and taking functions do not commute. What this inequality tells you is that, in general, you can not reason on the average.

So we're going to see instances where this property is not true. You're going to see lots of them. Let me just throw it here that it's something that's not true in general, but we will be interested in the exceptions where a relation like this is true. But these will be the exceptions. So in general, expectations are average, something like averages. But the function of an average is not the same as the average of the function.

OK. So now let's go to properties of expectations. Suppose that alpha is a real number, and I ask you, what's the expected value of that real number? So for example, if I write down this expression-- expected value of 2. What is this?

Well, we defined random variables and we defined expectations of random variables. So for this to make syntactic sense, this thing inside here should be a random variable. Is 2 -- the number 2 -- is it a random variable? In some sense, yes. It's the random variable that takes, always, the value of 2.

So suppose that you have some experiment and that experiment always outputs 2 whenever it happens. Then you can say, yes, it's a random experiment but it always gives me 2. The value of the random variable is always 2 no matter what. It's kind of a degenerate random variable that doesn't have any real randomness in it, but it's still useful to think of it as a special case.

So it corresponds to a function from the sample space to the real line that takes only one value. No matter what the outcome is, it always gives me a 2. OK. If you have a random variable that always gives you a 2, what is the expected value going to be? The only entry that shows up in this summation is that number 2. The probability of a 2 is equal to 1, and the value of that random variable is equal to 2. So it's the number itself. So the average value in an experiment that always gives you 2's is 2.

All right. So that's nice and simple. Now let's go to our experiment where age was your height in inches. And I know your height in inches, but I'm interested in your height measured in centimeters. How is that going to be related to your height in inches?

Well, if you take your height in inches and convert it to centimeters, I have another random variable, which is always, no matter what, two and a half times bigger than the random variable I started with. If you take some quantity and always multiplied by two and a half what happens to the average of that quantity? It also gets multiplied by two and a half. So you get a relation like this, which says that the average height of a student measured in centimeters is two and a half times the average height of a student measured in inches.

So that makes perfect intuitive sense. If you generalize it, it gives us this relation, that if you have a number, you can pull it outside the expectation and you get the right result. So this is a case where you can reason on the average. If you take a number, such as height, and multiply it by a certain number, you can reason on the average. I multiply the numbers by two, the averages will go up by two.

So this is an exception to this cautionary statement that I had up there. How do we prove that this fact is true? Well, we can use the expected value rule here, which tells us that the expected value of alpha X, this is our g of X , essentially, is going to be the sum over all x 's of my function, g of X , times the probability of the x 's. In our particular case, g of X is alpha times X . And we have those probabilities. And the alpha goes outside the summation. So we get alpha, sum over x 's, $x P_x$ of x , which is alpha times the expected value of X .

So that's how you prove this relation formally using this rule up here. And the next formula that I have here also gets proved the same way. What does this formula tell you? If I take everybody's height in centimeters-- we already multiplied by alpha-- and the gods give everyone a bonus of ten extra centimeters. What's going to happen to the average height of the class? Well, it will just go up by an extra ten centimeters.

So this expectation is going to be giving you the bonus of beta just adds a beta to the average height in centimeters, which we also know to be alpha times the expected value of X , plus beta. So this is a linearity property of expectations. If you take a linear function of a single random variable, the expected value of that linear function is the linear function of the expected value. So this is our big exception to this cautionary note, that we have equal if g is linear.

OK. All right. So let's get to the last concept of the day. What kind of functions of random variables may be of interest? One possibility might be the average value of X -squared. Why is it interesting? Well, why not.

It's the simplest function that you can think of. So if you want to calculate the expected value of X -squared, you would use this general rule for how you can calculate expected values of functions of random variables. You consider all the possible x 's. For each x , you see what's the probability that it occurs. And if that x occurs, you consider and see how big x -squared is.

Now, the more interesting quantity, a more interesting expectation that you can calculate has to do not with x -squared, but with the distance of x from the mean and then squared. So let's try to parse what we've got up here. Let's look just at the quantity inside here. What kind of quantity is it?

It's a random variable. Why? X is random, the random variable, expected value of X is a number. Subtract a number from a random variable, you get another random variable. Take a random variable and square it, you get another random variable. So the thing inside here is a legitimate random variable. What kind of random variable is it?

So suppose that we have our experiment and we have different x 's that can happen. And the mean of X in this picture might be somewhere around here. I do the experiment. I obtain some

numerical value of x . Let's say I obtain this numerical value. I look at the distance from the mean, which is this length, and I take the square of that.

Each time that I do the experiment, I go and record my distance from the mean and square it. So I give more emphasis to big distances. And then I take the average over all possible outcomes, all possible numerical values. So I'm trying to compute the average squared distance from the mean.

This corresponds to this formula here. So the picture that I drew corresponds to that. For every possible numerical value of x , that numerical value corresponds to a certain distance from the mean squared, and I weight according to how likely is that particular value of x to arise. So this measures the average squared distance from the mean.

Now, because of that expected value rule, of course, this thing is the same as that expectation. It's the average value of the random variable, which is the squared distance from the mean. With this probability, the random variable takes on this numerical value, and the squared distance from the mean ends up taking that particular numerical value.

OK. So why is the variance interesting? It tells us how far away from the mean we expect to be on the average. Well, actually we're not counting distances from the mean, it's distances squared. So it gives more emphasis to the kind of outliers in here. But it's a measure of how spread out the distribution is.

A big variance means that those bars go far to the left and to the right, typically. Where as a small variance would mean that all those bars are tightly concentrated around the mean value. It's the average squared deviation. Small variance means that we generally have small deviations. Large variances mean that we generally have large deviations.

Now as a practical matter, when you want to calculate the variance, there's a handy formula which I'm not proving but you will see it in recitation. It's just two lines of algebra. And it allows us to calculate it in a somewhat simpler way. We need to calculate the expected value of the random variable and the expected value of the squares of the random variable, and these two are going to give us the variance.

So to summarize what we did up here, the variance, by definition, is given by this formula. It's the expected value of the squared deviation. But we have the equivalent formula, which comes from application of the expected value rule, to the function g of X , equals to x minus the (expected value of X)-squared.

OK. So this is the definition. This comes from the expected value rule. What are some properties of the variance? Of course variances are always non-negative. Why is it always non-negative? Well, you look at the definition and you're just adding up non-negative things. We're adding squared deviations. So when you add non-negative things, you get something non-negative.

The next question is, how do things scale if you take a linear function of a random variable? Let's think about the effects of beta. If I take a random variable and add the constant to it, how does this affect the amount of spread that we have? It doesn't affect-- whatever the spread of this thing

is, if I add the constant beta, it just moves this diagram here, but the spread doesn't grow or get reduced.

The thing is that when I'm adding a constant to a random variable, all the x's that are going to appear are further to the right, but the expected value also moves to the right. And since we're only interested in distances from the mean, these distances do not get affected. x gets increased by something. The mean gets increased by that same something. The difference stays the same. So adding a constant to a random variable doesn't do anything to its variance.

But if I multiply a random variable by a constant alpha, what is that going to do to its variance? Because we have a square here, when I multiply my random variable by a constant, this x gets multiplied by a constant, the mean gets multiplied by a constant, the square gets multiplied by the square of that constant. And because of that reason, we get this square of alpha showing up here. So that's how variances transform under linear transformations. You multiply your random variable by constant, the variance goes up by the square of that same constant. OK. That's it for today. See you on Wednesday.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013 Transcript – Tutorial: PMF of a Function of a Random Variable

Hey guys. Welcome back. Today, we're going to be working on a problem that asks you to find the PMF of a function of a random variable. So let's just jump right in. The problem statement gives you the PMF for a random variable called x . So we're told that there's this random variable x that takes on values minus 3, minus 2, minus 1, 1, 2, and 3.

And for each of those values, the probability mass lying over that value is given by this formula, x squared over a . Now I didn't write it here to save room, but we're also told that a is a real number that is greater than 0. And we're told that the probability of x taking on any value outside of the set is 0.

Now we're asked to do two things in the problem. First is to find the value of the parameter a . And that's sort of a natural question to ask, because if you think about it, the PMF isn't fully specified. And in fact, if you plug in the wrong number for a , you actually won't get a valid PMF. So we'll explore that idea in the first part.

And then the second part, you're given a new random variable called z . And z happens to be a function of x . In fact, it's equal to x squared. And then you're asked to compute that PMF. So this problem is a good practice problem. I think, at this point, you guys are sort of newly acquainted with the idea of a PMF, or probability mass function. So this problem will hopefully help you get more familiar with that concept and how to manipulate PMFs.

And by the way, just to make sure we're all on the same page, what does a PMF really tell you? So $p_{\text{sub } X}$, where this is a capital X , because the convention in this class is to use capital letters for random variables. So p_X of k , this is defined to be the probability that your random variable X takes on a value of k .

So essentially, this says-- and this is just some number. So in our particular case, this would be equal to k squared over a . And how you can interpret this is this p_X guy is sort of like a machine. He takes in some value that your random variable could take on, and then he spits out the amount of probability mass lying over that value. OK.

So now that we've done that quick recap, let's get back to the first part of the problem. So we have this formula for p_X of x , and we need to solve for a . So in order to do that, we're going to use one of our axioms of probability to set up an equation. And then we can solve precisely for a . So namely, we know that every PMF must sum to 1. And so essentially, if you sum this guy over all possible values of x , you should get a 1, and that equation will let us solve for a .

So let's do that. Summation over x of p_X of x . So here, essentially you're only summing over these six values. So this is equal to p_X of minus 3, plus p_X of minus 2, plus p_X of minus 1, et

cetera. Oops. p_x of 2 plus p_x of 3. OK. And again, like the interpretation as we said, this number here should be interpreted as the amount of probability mass lying over minus 3.

And to help you visualize this, actually, before we go further with the computation, let's actually plot this PMF. So the amount of probability mass lying over minus 3, the way we figure that out is we take minus 3 and we plug it into this formula up here. So you get $9/a$. Now you can do this for minus 2. You've got $4/a$, looking at the formula. For 1, you get $1/a$. And of course, this graph, it's the mirror image over 0, because of the symmetry.

So hopefully this little visualization helps you understand what I'm talking about. And now we can just read these values off of the plot we just made. So we know p_x minus 3 is equal to p_x of 3. So we can go ahead and just take 2 times $9/a$. Similarly, we get 2 times $4/a$, and then plus 2 times $1/a$. So now it's just a question of algebra.

So simplifying this, you're going to get 18 plus 8 plus 2, divided by a . And this gives you $28/a$. And as I argued before, you know that if you sum a PMF over all possible values, you must get 1. So this is equal to 1, which of course implies that a is equal to 28.

So what we've shown here is that you actually don't have a choice for what value a can take on. It must take on 28. And in fact, if you plug in any other value than 28 in here, you actually are not going to have a valid PMF, because it's not going to sum to 1. OK. So I'm going to write my answer here, and then erase to give myself more room for part

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013

Transcript – Recitation: Sampling People on Buses

Hi. In this problem, we're dealing with buses of students going to a job convention. And in the problem, we'll be exercising our knowledge of PMFs-- probability mass functions. So we'll get a couple of opportunities to write out some PMFs, and also calculating expectations or expected values. And also, importantly, we'll actually be exercising our intuition to help us not just rely on numbers, but also to just have a sense of what the answers to some probability questions should be.

So the problem specifically deals with four buses of students. So we have buses, and in each one carries a different number of students. So the first one carries 40 students, the second one 33, the third one has 25, and the last one has 50 students for a total of 148 students. And because these students are smart, and they like probability, they are interested in a couple questions.

So suppose that one of these 148 students is chosen randomly, and so we'll assume that what that means is that each one has the same probability of being chosen. So they're chosen uniformly at random. And let's assign a couple of random variables. So we'll say x corresponds to the number of students in the bus of the selected student.

OK, so one of these 148 students is selected uniformly at random, and we'll let x correspond to the number of students in that student's bus. So if a student from this bus was chosen, then x would be 25, for example.

OK, and then let's come up with another random variable, y , which is almost the same thing. Except instead of now selecting a random student, we'll select a random bus. Or equivalently, we'll select a random bus driver.

So each bus has one driver, and instead of selecting one of the 148 students at random, we'll select one of the four bus drivers also uniformly at random. And we'll say the number of students in that driver's bus will be y . So for example, if this bus driver was selected, then y would be 33.

OK, so the main problem that we're trying to answer is what do you expect the expectation-- which one of these random variables do you expect to have the higher expectation or the higher expected value? So, would you expect x to be higher on average, or y to be higher? And what would be the intuition for this?

So obviously, we can actually write out the PMFs for x and y . These are just discrete random variables. And we can actually calculate out what the expectation is. But it's also useful to exercise your intuition, and your sense of what the answer should be.

So it might not be immediately clear which one would be higher, or you might even say that maybe it doesn't make a difference. They're actually the same. But a useful way to approach some of these questions is to try to take things to the extreme and see how that plays out.

So let's take the simpler example and take it to the extreme and say, suppose a set of four buses carrying these number of students. We have only two buses-- one bus that has only 1 student, and we have another bus that has 1,000 students. OK. And suppose we ask the same question.

Well, now if you look at it, there's a total of 1,001 students now. If you select one of the students at random, it's overwhelmingly more likely that that student will be one of the 1,000 students on this huge bus. It's very unlikely that you'll get lucky and select the one student who is by himself.

And so because of that, you have a very high chance of selecting the bus with the high number of students. And so you would expect x , the number of students, to be high-- to be almost 1,000 in the expectation. But on the other hand, if you selected the driver at random, then you have a 50/50 chance of selecting this one or that one. And so you would expect the expectation there to be roughly 500 or so. And so you can see that if you take this to the extreme, then it becomes more clear what the answer would be.

And the argument is that the expectation of x should be higher than the expectation of y , and the reason here is that because you select the student at random, you're more likely to select a student who is in a large bus, because that bus just has more students to select from. And because of that, you're more biased in favor of selecting large buses, and therefore, that makes x higher in expectation. OK, so that's the intuition behind this problem. And now, as I actually go through some of the more mechanics and write out what the PMFs and the calculation for the expectation would be to verify that our intuition is actually correct.

OK, so we have two random variables that are defined. Now let's just write out what their PMFs are. So the PMF-- we write it as little P of capital X and little x . So the random variable-- what we do is we say the probability that it will take on a certain value, right? So what is the probability that x will be 40?

Well, x will be 40 if a student from this bus was selected. And what's the probability that a student from this bus is selected? That probability is 40/148, because there's 148 students, 40 of whom are sitting in this bus. And similarly, x will be 33 with probability 33/148, and x will be 25 with probability 25/148. And x will be 50 with probability 50/148. And it will be 0 otherwise.

OK, so there is our PMF for x , and we can do the same thing for y . The PMF of y -- again, we say what is the probability that y will take on certain values? Well, y can take on the same values as x can, because we're still dealing with the number of students in each bus. So y can be 40.

But the probability that y is 40, because we're selecting the driver at random now, is 1/4, right? Because there's a 1/4 chance that we'll pick this driver. And the probability that y will be 33 will also be 1/4, and the same thing for 25 and 50. And it's 0 otherwise.

OK, so those are the PMFs for our two random variables, x and y . And we can also draw out what the PMFs look like. So if this is 25, 30, 35, 40, 45, and 50, then the probability that it's 25 is 25/148. So we can draw a mass right there.

For 33, it's a little higher, because it's 33/148 instead of 25. For 40, it's even higher still. It's 40/148. And for 50, it is still higher, because it is 50/148. And so you can see that the PMF is more heavily favored towards the larger values.

We can do the same thing for y , and we'll notice that there's a difference in how these distributions look. So if we do the same thing, the difference now is that all four of these masses will have the same height. Each one will have height 1/4, whereas this one for x , it's more heavily biased in favor of the larger ones. And so because of that, we can actually now calculate what the expectations are and figure out whether or not our intuition was correct.

OK, so now let's actually calculate out what these expectations are. So as you recall, the expectation is calculated out as a weighted sum. So for each possible value of x , you take that value and you weight it by the probability of the random variable taking on that value. So in this case, it would be 40 times 40/148, 33 times 33/148, and so on.

48 plus 25 times 25/148 plus 50 times 50/148. And if you do out this calculation, what you'll get is that it is around 39. Roughly 39.

And now we can do the same thing for y . But for y , it's different, because now instead of weighting it by these probabilities, we'll weight it by these probabilities. So each one has the same weight of 1/4.

So now we get 40 times 1/4 plus 33 times 1/4. That's 25 times 1/4 plus 50 times 1/4. And if you do out this arithmetic, what you get is that this expectation is 37. And so what we get is that, in fact, after we do out the calculations, the expected value of x is indeed greater than the expected value of y , which confirms our intuition.

OK, so this problem, to summarize-- we've reviewed how to write out a PMF and also how to calculate expectations. But also, we've got a chance to figure out some intuition behind some of these problems. And so sometimes it's helpful to take simpler things and take things to the extreme and figure out intuitively whether or not the answer makes sense.

It's useful just to verify whether the numerical answer that you get in the end is correct. Does this actually make sense? It's a useful guide for when you're solving these problems. OK, so we'll see you next time.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Problem Set 4
Due October 6, 2010

1. Random variables X and Y have the joint PMF

$$p_{X,Y}(x,y) = \begin{cases} c(x^2 + y^2), & \text{if } x \in \{1, 2, 4\} \text{ and } y \in \{1, 3\}, \\ 0, & \text{otherwise.} \end{cases}$$

- (a) What is the value of the constant c ?
 - (b) What is $\mathbf{P}(Y < X)$?
 - (c) What is $\mathbf{P}(Y > X)$?
 - (d) What is $\mathbf{P}(Y = X)$?
 - (e) What is $\mathbf{P}(Y = 3)$?
 - (f) Find the marginal PMFs $p_X(x)$ and $p_Y(y)$.
 - (g) Find the expectations $\mathbf{E}[X]$, $\mathbf{E}[Y]$ and $\mathbf{E}[XY]$.
 - (h) Find the variances $\text{var}(X)$, $\text{var}(Y)$ and $\text{var}(X + Y)$.
 - (i) Let A denote the event $X \geq Y$. Find $\mathbf{E}[X | A]$ and $\text{var}(X | A)$.
2. The newest invention of the 6.041/6.431 staff is a three-sided die with faces numbered 1, 2, and 3. The PMF for the result of any one roll of this die is

$$p_X(x) = \begin{cases} 1/2, & \text{if } x = 1, \\ 1/4, & \text{if } x = 2, \\ 1/4, & \text{if } x = 3, \\ 0, & \text{otherwise.} \end{cases}$$

Consider a sequence of six independent rolls of this die, and let X_i be the random variable corresponding to the i th roll.

- (a) What is the probability that exactly three of the rolls have result equal to 3?
 - (b) What is the probability that the first roll is 1, given that exactly two of the six rolls have result of 1?
 - (c) We are told that exactly three of the rolls resulted in 1 and exactly three resulted in 2. Given this information, what is the probability that the sequence of rolls is 121212?
 - (d) Conditioned on the event that at least one roll resulted in 3, find the conditional PMF of the number of 3's.
3. Suppose that X and Y are independent, identically distributed, geometric random variables with parameter p . Show that

$$\mathbf{P}(X = i | X + Y = n) = \frac{1}{n-1}, \quad \text{for } i = 1, 2, \dots, n-1.$$

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Fall 2010)

4. Consider 10 independent tosses of a biased coin with a probability of heads of p .
 - (a) Let A be the event that there are 6 heads in the first 8 tosses. Let B be the event that the 9th toss results in heads. Show that events A and B are independent.
 - (b) Find the probability that there are 3 heads in the first 4 tosses and 2 heads in the last 3 tosses.
 - (c) Given that there were 4 heads in the first 7 tosses, find the probability that the 2nd head occurred during the 4th trial.
 - (d) Find the probability that there are 5 heads in the first 8 tosses and 3 heads in the last 5 tosses.
5. Consider a sequence of independent tosses of a biased coin at times $t = 0, 1, 2, \dots$. On each toss, the probability of a 'head' is p , and the probability of a 'tail' is $1 - p$. A reward of one unit is given each time that a 'tail' follows immediately after a 'head.' Let R be the total reward paid in times $1, 2, \dots, n$. Find $\mathbf{E}[R]$ and $\text{var}(R)$.

G1[†]. A simple example of a random variable is the *indicator* of an event A , which is denoted by I_A :

$$I_A(\omega) = \begin{cases} 1, & \text{if } \omega \in A \\ 0, & \text{otherwise.} \end{cases}$$

- (a) Prove that two events A and B are independent if and only if the associated indicator random variables, I_A and I_B are independent.
- (b) Show that if $X = I_A$, then $\mathbf{E}[X] = \mathbf{P}(A)$.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Problem Set 4: Solutions

1. (a) From the joint PMF, there are six (x, y) coordinate pairs with nonzero probabilities of occurring. These pairs are $(1, 1)$, $(1, 3)$, $(2, 1)$, $(2, 3)$, $(4, 1)$, and $(4, 3)$. The probability of a pair is proportional to the sum of the squares of the coordinates of the pair, $x^2 + y^2$. Because the probability of the entire sample space must equal 1, we have:

$$(1+1)c + (1+9)c + (4+1)c + (4+9)c + (16+1)c + (16+9)c = 1.$$

Solving for c , we get $c = \boxed{\frac{1}{72}}$.

- (b) There are three sample points for which $y < x$:

$$\mathbf{P}(Y < X) = \mathbf{P}(\{(2, 1)\}) + \mathbf{P}(\{(4, 1)\}) + \mathbf{P}(\{(4, 3)\}) = \frac{5}{72} + \frac{17}{72} + \frac{25}{72} = \boxed{\frac{47}{72}}.$$

- (c) There are two sample points for which $y > x$:

$$\mathbf{P}(Y > X) = \mathbf{P}(\{(1, 3)\}) + \mathbf{P}(\{(2, 3)\}) = \frac{10}{72} + \frac{13}{72} = \boxed{\frac{23}{72}}.$$

- (d) There is only one sample point for which $y = x$:

$$\mathbf{P}(Y = X) = \mathbf{P}(\{(1, 1)\}) = \boxed{\frac{2}{72}}.$$

Notice that, using the above two parts,

$$\mathbf{P}(Y < X) + \mathbf{P}(Y > X) + \mathbf{P}(Y = X) = \frac{47}{72} + \frac{23}{72} + \frac{2}{72} = 1$$

as expected.

- (e) There are three sample points for which $y = 3$:

$$\mathbf{P}(Y = 3) = \mathbf{P}(\{(1, 3)\}) + \mathbf{P}(\{(2, 3)\}) + \mathbf{P}(\{(4, 3)\}) = \frac{10}{72} + \frac{13}{72} + \frac{25}{72} = \boxed{\frac{48}{72}}.$$

- (f) In general, for two discrete random variable X and Y for which a joint PMF is defined, we have

$$p_X(x) = \sum_{y=-\infty}^{\infty} p_{X,Y}(x, y) \quad \text{and} \quad p_Y(y) = \sum_{x=-\infty}^{\infty} p_{X,Y}(x, y).$$

In this problem the ranges of X and Y are quite restricted so we can determine the marginal PMFs by enumeration. For example,

$$p_X(2) = \mathbf{P}(\{(2, 1)\}) + \mathbf{P}(\{(2, 3)\}) = \frac{18}{72}.$$

Overall, we get:

$$p_X(x) = \begin{cases} 12/72, & \text{if } x = 1, \\ 18/72, & \text{if } x = 2, \\ 42/72, & \text{if } x = 4, \\ 0, & \text{otherwise} \end{cases} \quad \text{and} \quad p_Y(y) = \begin{cases} 24/72, & \text{if } y = 1, \\ 48/72, & \text{if } y = 3, \\ 0, & \text{otherwise.} \end{cases}$$

- (g) In general, the expected value of any discrete random variable X equals

$$\mathbf{E}[X] = \sum_{x=-\infty}^{\infty} xp_X(x).$$

For this problem,

$$\mathbf{E}[X] = 1 \cdot \frac{12}{72} + 2 \cdot \frac{18}{72} + 4 \cdot \frac{42}{72} = \boxed{3}$$

and

$$\mathbf{E}[Y] = 1 \cdot \frac{24}{72} + 3 \cdot \frac{48}{72} = \boxed{\frac{7}{3}}.$$

To compute $\mathbf{E}[XY]$, note that $p_{X,Y}(x,y) \neq p_X(x)p_Y(y)$. Therefore, X and Y are not independent and we cannot assume $\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y]$. Thus, we have

$$\begin{aligned}\mathbf{E}[XY] &= \sum_x \sum_y xy p_{X,Y}(x,y) \\ &= 1 \cdot \frac{2}{72} + 2 \cdot \frac{5}{72} + 4 \cdot \frac{17}{72} + 3 \cdot \frac{10}{72} + 6 \cdot \frac{13}{72} + 12 \cdot \frac{25}{72} = \boxed{\frac{61}{9}}.\end{aligned}$$

- (h) The variance of a random variable X can be computed as $\mathbf{E}[X^2] - \mathbf{E}[X]^2$ or as $\mathbf{E}[(X - \mathbf{E}[X])^2]$.

We use the second approach here because X and Y take on such limited ranges. We have

$$\text{var}(X) = (1-3)^2 \frac{12}{72} + (2-3)^2 \frac{18}{72} + (4-3)^2 \frac{42}{72} = \boxed{\frac{3}{2}}$$

and

$$\text{var}(Y) = (1-\frac{7}{3})^2 \frac{24}{72} + (3-\frac{7}{3})^2 \frac{48}{72} = \boxed{\frac{8}{9}}.$$

X and Y are not independent, so we cannot assume $\text{var}(X+Y) = \text{var}(X) + \text{var}(Y)$. The variance of $X+Y$ will be computed using $\text{var}(X+Y) = \mathbf{E}[(X+Y)^2] - (\mathbf{E}[X+Y])^2$. Therefore, we have

$$\begin{aligned}\mathbf{E}[(X+Y)^2] &= 4 \cdot \frac{2}{72} + 9 \cdot \frac{5}{72} + 25 \cdot \frac{17}{72} + 16 \cdot \frac{10}{72} + 25 \cdot \frac{13}{72} + 49 \cdot \frac{25}{72} = \frac{547}{18}. \\ (\mathbf{E}[X+Y])^2 &= (\mathbf{E}[X] + \mathbf{E}[Y])^2 = \left(3 + \frac{7}{3}\right)^2 = \frac{256}{9}.\end{aligned}$$

Therefore,

$$\text{var}(X+Y) = \frac{547}{18} - \frac{256}{9} = \boxed{\frac{35}{18}}.$$

- (i) There are four (x, y) coordinate pairs in A : $(1,1)$, $(2,1)$, $(4,1)$, and $(4,3)$. Therefore, $\mathbf{P}(A) = \frac{1}{72}(2 + 5 + 17 + 25) = \frac{49}{72}$. To find $\mathbf{E}[X | A]$ and $\text{var}(X | A)$, $p_{X|A}(x)$ must be calculated. We have

$$p_{X|A}(x) = \begin{cases} 2/49, & \text{if } x = 1, \\ 5/49, & \text{if } x = 2, \\ 42/49, & \text{if } x = 4, \\ 0, & \text{otherwise,} \end{cases}$$

$$\begin{aligned}\mathbf{E}[X | A] &= 1 \cdot \frac{2}{49} + 2 \cdot \frac{5}{49} + 4 \cdot \frac{42}{49} = \boxed{\frac{180}{49}}, \\ \mathbf{E}[X^2 | A] &= 1^2 \cdot \frac{2}{49} + 2^2 \cdot \frac{5}{49} + 4^2 \cdot \frac{42}{49} = \frac{694}{49}, \\ \text{var}(X | A) &= \mathbf{E}[X^2 | A] - (\mathbf{E}[X | A])^2 = \frac{694}{49} - \left(\frac{180}{49}\right)^2 = \boxed{\frac{1606}{2401}},\end{aligned}$$

2. Consider a sequence of six independent rolls of this die, and let X_i be the random variable corresponding to the i th roll.

- (a) What is the probability that exactly three of the rolls have result equal to 3? Each roll X_i can either be a 3 with probability $1/4$ or not a 3 with probability $3/4$. There are $\binom{6}{3}$ ways of placing the 3's in the sequence of six rolls. After we require that a 3 go in each of these spots, which has probability $(1/4)^3$, our only remaining condition is that either a 1 or a 2 go in the other three spots, which has probability $(3/4)^3$. So the probability of exactly three rolls of 3 in a sequence of six independent rolls is $\boxed{\binom{6}{3}(\frac{1}{4})^3(\frac{3}{4})^3}$.
- (b) What is the probability that the first roll is 1, given that exactly two of the six rolls have result of 1? The probability of obtaining a 1 on a single roll is $1/2$, and the probability of obtaining a 2 or 3 on a single roll is also $1/2$. For the purposes of solving this problem we treat obtaining a 2 or 3 as an equivalent result. We know that there are $\binom{6}{2}$ ways of rolling exactly two 1's. Of these $\binom{6}{2}$ ways, exactly $\binom{5}{1} = 5$ ways result in a 1 in the first roll, since we can place the remaining 1 in any of the five remaining rolls. The rest of the rolls must be either 2 or 3. Thus, the probability that the first roll is a 1 given exactly two rolls had an outcome of 1 is $\boxed{\frac{5}{\binom{6}{2}}}$.
- (c) We are now told that exactly three of the rolls resulted in 1 and exactly three resulted in 2. What is the probability of the sequence 121212? We want to find

$$\mathbf{P}(121212 | \text{exactly three 1's and three 2's}) = \frac{\mathbf{P}(121212)}{\mathbf{P}(\text{exactly 3 ones and 3 twos})}.$$

Any particular sequence of three 1's and three 2's will have the same probability: $(1/2)^3(1/4)^3$. There are $\binom{6}{3}$ possible rolls with exactly three 1's and three 2's. Therefore,

$$\mathbf{P}(121212 | \text{exactly three 1's and three 2's}) = \boxed{\frac{1}{\binom{6}{3}}}.$$

- (d) Conditioned on the event that at least one roll resulted in 3, find the conditional PMF of the number of 3's. Let A be the event that at least one roll results in a 3. Then

$$\mathbf{P}(A) = 1 - \mathbf{P}(\text{no rolls resulted in 3}) = 1 - \left(\frac{3}{4}\right)^6.$$

Now let K be the random variable representing the number of 3's in the 6 rolls. The (unconditional) PMF $p_K(k)$ for K is given by

$$p_K(k) = \binom{6}{k} \left(\frac{1}{4}\right)^k \left(\frac{3}{4}\right)^{6-k}.$$

We find the conditional PMF $p_{K|A}(k | A)$ using the definition of conditional probability:

$$p_{K|A}(k | A) = \frac{\mathbf{P}(\{K = k\} \cap A)}{\mathbf{P}(A)}.$$

Thus we obtain

$$p_{K|A}(k | A) = \begin{cases} \frac{1}{1-(3/4)^6} \binom{6}{k} \left(\frac{1}{4}\right)^k \left(\frac{3}{4}\right)^{6-k} & \text{if } k = 1, 2, \dots, 6, \\ 0 & \text{otherwise.} \end{cases}$$

Note that $p_{K|A}(0 | A) = 0$ because the event $\{K = 0\}$ and the event A are mutually exclusive. Thus the probability of their intersection, which appears in the numerator in the definition of the conditional PMF, is zero.

3. By the definition of conditional probability,

$$\mathbf{P}(X = i | X + Y = n) = \frac{\mathbf{P}(\{X = i\} \cap \{X + Y = n\})}{\mathbf{P}(X + Y = n)}.$$

The event $\{X = i\} \cap \{X + Y = n\}$ in the numerator is equivalent to $\{X = i\} \cap \{Y = n - i\}$. Combining this with the independence of X and Y ,

$$\mathbf{P}(\{X = i\} \cap \{X + Y = n\}) = \mathbf{P}(\{X = i\} \cap \{Y = n - i\}) = \mathbf{P}(X = i)\mathbf{P}(Y = n - i).$$

In the denominator, $\mathbf{P}(X + Y = n)$ can be expanded using the total probability theorem and the independence of X and Y :

$$\begin{aligned} \mathbf{P}(X + Y = n) &= \sum_{i=1}^{n-1} \mathbf{P}(X = i)\mathbf{P}(X + Y = n | X = i) \\ &= \sum_{i=1}^{n-1} \mathbf{P}(X = i)\mathbf{P}(i + Y = n | X = i) \\ &= \sum_{i=1}^{n-1} \mathbf{P}(X = i)\mathbf{P}(Y = n - i | X = i) \\ &= \sum_{i=1}^{n-1} \mathbf{P}(X = i)\mathbf{P}(Y = n - i) \end{aligned}$$

Note that we only get non-zero probability for $i = 1, \dots, n - 1$ since X and Y are geometric random variables.

The desired result is obtained by combining the computations above and using the geometric

PMF explicitly:

$$\begin{aligned}
 \mathbf{P}(X = i \mid X + Y = n) &= \frac{\mathbf{P}(X = i)\mathbf{P}(Y = n - i)}{\sum_{i=1}^{n-1} \mathbf{P}(X = i)\mathbf{P}(Y = n - i)} \\
 &= \frac{(1-p)^{i-1}p(1-p)^{n-i-1}p}{\sum_{i=1}^{n-1} (1-p)^{i-1}p(1-p)^{n-i-1}p} \\
 &= \frac{(1-p)^n}{\sum_{i=1}^{n-1} (1-p)^n} \\
 &= \frac{(1-p)^n}{(1-p)^n \sum_{i=1}^{n-1} 1} \\
 &= \frac{1}{n-1}, \quad i = 1, \dots, n-1.
 \end{aligned}$$

4. (a) Since $\mathbf{P}(A) > 0$, we can show independence through $\mathbf{P}(B) = \mathbf{P}(B \mid A)$:

$$\mathbf{P}(B \mid A) = \frac{\mathbf{P}(B \cap A)}{\mathbf{P}(A)} = \frac{\binom{8}{6}p^6(1-p)^2p}{\binom{8}{6}p^6(1-p)^2} = p = \mathbf{P}(B).$$

Therefore, A and B are independent.

- (b) Let C be the event “3 heads in the first 4 tosses” and let D be the event “2 heads in the last 3 tosses”. Since there are no overlap in tosses in C and D , they are independent:

$$\begin{aligned}
 \mathbf{P}(C \cap D) &= \mathbf{P}(C)\mathbf{P}(D) \\
 &= \binom{4}{3}p^3(1-p) \cdot \binom{3}{2}p^2(1-p) \\
 &= 12p^5(1-p)^2.
 \end{aligned}$$

- (c) Let E be the event “4 heads in the first 7 tosses” and let F be the event “2nd head occurred during 4th trial”. We are asked to find $\mathbf{P}(F \mid E) = \mathbf{P}(F \cap E)/\mathbf{P}(E)$. The event $F \cap E$ occurs if there is 1 head in the first 3 trials, 1 head on the 4th trial, and 2 heads in the last 3 trials. Thus, we have

$$\begin{aligned}
 \mathbf{P}(F \mid E) &= \frac{\mathbf{P}(F \cap E)}{\mathbf{P}(E)} = \frac{\binom{3}{1}p(1-p)^2 \cdot p \cdot \binom{3}{2}p^2(1-p)}{\binom{7}{4}p^4(1-p)^3} \\
 &= \frac{\binom{3}{1} \cdot 1 \cdot \binom{3}{2}}{\binom{7}{4}} = \frac{9}{35}.
 \end{aligned}$$

Alternatively, we can solve this by counting. We are given that 4 heads occurred in the first 7 tosses. Each sequence of 7 trials with 4 heads is equally probable, the discrete uniform

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
 (Fall 2010)

probability law can be used here. There are $\binom{7}{4}$ outcomes in E . For the event $E \cap F$, there are $\binom{3}{1}$ ways to arrange 1 head in the first 3 trials, 1 way to arrange the 2nd head in the 4th trial and $\binom{3}{2}$ ways to arrange 2 heads in the first 3 trials. Therefore,

$$\mathbf{P}(F | E) = \frac{\binom{3}{1} \cdot 1 \cdot \binom{3}{2}}{\binom{7}{4}} = \frac{9}{35}.$$

- (d) Let G be the event “5 heads in the first 8 tosses” and let H be the event “3 heads in the last 5 tosses”. These two events are not independent as there is some overlap in the tosses (the 6th, 7th, and 8th tosses). To compute the probability of interest, we carefully count all the disjoint, possible outcomes in the set $G \cap H$ by conditioning on the number of heads in the 6th, 7th, and the 8th tosses. We have

$$\begin{aligned}\mathbf{P}(G \cap H) &= \mathbf{P}(G \cap H | 1 \text{ head in tosses 6--8})\mathbf{P}(1 \text{ head in tosses 6--8}) \\ &\quad + \mathbf{P}(G \cap H | 2 \text{ heads in tosses 6--8})\mathbf{P}(2 \text{ heads in tosses 6--8}) \\ &\quad + \mathbf{P}(G \cap H | 3 \text{ heads in tosses 6--8})\mathbf{P}(3 \text{ heads in tosses 6--8}) \\ &= \binom{5}{4}p^4(1-p) \cdot p^2 \cdot \binom{3}{1}p(1-p)^2 \\ &\quad + \binom{5}{3}p^3(1-p)^2 \cdot \binom{2}{1}p(1-p) \cdot \binom{3}{2}p^2(1-p) \\ &\quad + \binom{5}{2}p^2(1-p)^3 \cdot (1-p)^2 \cdot p^3. \\ &= 15p^7(1-p)^3 + 60p^6(1-p)^4 + 10p^5(1-p)^5.\end{aligned}$$

5. Let I_k be the reward paid at time k . We have

$$\mathbf{E}[I_k] = \mathbf{P}(I_k = 1) = \mathbf{P}(\text{T at time } k \text{ and H at time } k-1) = p(1-p).$$

Computing $\mathbf{E}[R]$ is immediate because

$$\mathbf{E}[R] = \mathbf{E} \left[\sum_{k=1}^n I_k \right] = \sum_{k=1}^n \mathbf{E}[I_k] = np(1-p).$$

The variance calculation is not as easy because the I_k s are not all independent:

$$\begin{aligned}\mathbf{E}[I_k^2] &= p(1-p) \\ \mathbf{E}[I_k I_{k+1}] &= 0 \quad \text{because rewards at times } k \text{ and } k+1 \text{ are inconsistent} \\ \mathbf{E}[I_k I_{k+\ell}] &= \mathbf{E}[I_k]\mathbf{E}[I_{k+\ell}] = p^2(1-p)^2 \quad \text{for } \ell \geq 2\end{aligned}$$

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
 (Fall 2010)

$$\begin{aligned}
 \mathbf{E}[R^2] &= \mathbf{E}\left[\left(\sum_{k=1}^n I_k\right)\left(\sum_{m=1}^n I_m\right)\right] = \sum_{k=1}^n \sum_{m=1}^n \mathbf{E}[I_k I_m] \\
 &= \underbrace{np(1-p)}_{n \text{ terms with } k=m} + \underbrace{0}_{2(n-1) \text{ terms with } |k-m|=1} + \underbrace{(n^2 - 3n + 2)p^2(1-p)^2}_{n^2 - 3n + 2 \text{ terms with } |k-m|>1} \\
 \text{var}(R) &= \mathbf{E}[R^2] - (\mathbf{E}[R])^2 \\
 &= np(1-p) + (n^2 - 3n + 2)p^2(1-p)^2 - n^2 p^2(1-p)^2 \\
 &= np(1-p) - (3n - 2)p^2(1-p)^2.
 \end{aligned}$$

- G1[†]. (a) We know that I_A is a random variable that maps a 1 to the real number line if ω occurs within an event A and maps a 0 to the real number line if ω occurs outside of event A . A similar argument holds for event B . Thus we have,

$$I_A(\omega) = \begin{cases} 1, & \text{with probability } \mathbf{P}(A) \\ 0, & \text{with probability } 1 - \mathbf{P}(A) \end{cases}$$

$$I_B(\omega) = \begin{cases} 1, & \text{with probability } \mathbf{P}(B) \\ 0, & \text{with probability } 1 - \mathbf{P}(B) \end{cases}$$

If the random variables, A and B , are independent, we have $\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B)$. The indicator random variables, I_A and I_B , are independent if, $\mathbf{P}_{I_A, I_B}(x, y) = \mathbf{P}_{I_A}(x)\mathbf{P}_{I_B}(y)$

We know that the intersection of A and B yields.

$$\begin{aligned}
 \mathbf{P}_{I_A, I_B}(1, 1) &= \mathbf{P}_{I_A}(1)\mathbf{P}_{I_B}(1) \\
 &= \mathbf{P}(A)\mathbf{P}(B) \\
 &= \mathbf{P}(A \cap B)
 \end{aligned}$$

We also have,

$$\begin{aligned}
 \mathbf{P}_{I_A, I_B}(1, 1) &= \mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B) = \mathbf{P}_{I_A}(1)\mathbf{P}_{I_B}(1) \\
 \mathbf{P}_{I_A, I_B}(0, 1) &= \mathbf{P}(A^c \cap B) = \mathbf{P}(A^c)\mathbf{P}(B) = \mathbf{P}_{I_A}(0)\mathbf{P}_{I_B}(1) \\
 \mathbf{P}_{I_A, I_B}(1, 0) &= \mathbf{P}(A \cap B^c) = \mathbf{P}(A)\mathbf{P}(B^c) = \mathbf{P}_{I_A}(1)\mathbf{P}_{I_B}(0) \\
 \mathbf{P}_{I_A, I_B}(0, 0) &= \mathbf{P}(A^c \cap B^c) = \mathbf{P}(A^c)\mathbf{P}(B^c) = \mathbf{P}_{I_A}(0)\mathbf{P}_{I_B}(0)
 \end{aligned}$$

- (b) If $X = I_A$, we know that

$$\mathbf{E}[X] = \mathbf{E}[I_A] = 1 \cdot \mathbf{P}(A) + 0 \cdot (1 - \mathbf{P}(A)) = \mathbf{P}(A)$$

[†]Required for 6.431; optional for 6.041

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

LECTURE 5

- **Readings:** Sections 2.1-2.3, start 2.4

Lecture outline

- Random variables
- Probability mass function (PMF)
- Expectation
- Variance

Random variables

- An assignment of a value (number) to every possible outcome
- Mathematically: A function from the sample space Ω to the real numbers
 - discrete or continuous values
- Can have several random variables defined on the same sample space
- Notation:
 - random variable X
 - numerical value x

Probability mass function (PMF)

- (“probability law”, “probability distribution” of X)
- Notation:

$$\begin{aligned} p_X(x) &= \mathbf{P}(X = x) \\ &= \mathbf{P}(\{\omega \in \Omega \text{ s.t. } X(\omega) = x\}) \end{aligned}$$

- $p_X(x) \geq 0 \quad \sum_x p_X(x) = 1$

- **Example:** X =number of coin tosses until first head

– assume independent tosses,
 $\mathbf{P}(H) = p > 0$

$$\begin{aligned} p_X(k) &= \mathbf{P}(X = k) \\ &= \mathbf{P}(TT \cdots TH) \\ &= (1-p)^{k-1}p, \quad k = 1, 2, \dots \end{aligned}$$

– **geometric PMF**

How to compute a PMF $p_X(x)$

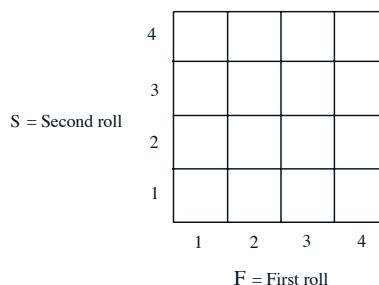
- collect all possible outcomes for which X is equal to x
- add their probabilities
- repeat for all x

- **Example:** Two independent rolls of a fair tetrahedral die

F : outcome of first throw

S : outcome of second throw

$$X = \min(F, S)$$



$$p_X(2) =$$

Binomial PMF

- X : number of heads in n independent coin tosses

- $P(H) = p$

- Let $n = 4$

$$\begin{aligned} p_X(2) &= P(HHTT) + P(HTHT) + P(HTTH) \\ &\quad + P(THHT) + P(THTH) + P(TTHH) \\ &= 6p^2(1-p)^2 \\ &= \binom{4}{2}p^2(1-p)^2 \end{aligned}$$

In general:

$$p_X(k) = \binom{n}{k}p^k(1-p)^{n-k}, \quad k = 0, 1, \dots, n$$

Expectation

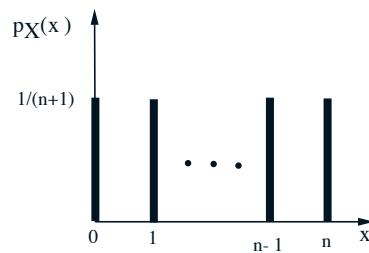
- Definition:

$$E[X] = \sum_x xp_X(x)$$

- Interpretations:

- Center of gravity of PMF
- Average in large number of repetitions of the experiment
(to be substantiated later in this course)

- Example: Uniform on $0, 1, \dots, n$



$$E[X] = 0 \times \frac{1}{n+1} + 1 \times \frac{1}{n+1} + \dots + n \times \frac{1}{n+1} =$$

Properties of expectations

- Let X be a r.v. and let $Y = g(X)$

- Hard: $E[Y] = \sum_y y p_Y(y)$

- Easy: $E[Y] = \sum_x g(x) p_X(x)$

- Caution: In general, $E[g(X)] \neq g(E[X])$

Properties: If α, β are constants, then:

- $E[\alpha] =$

- $E[\alpha X] =$

- $E[\alpha X + \beta] =$

Variance

Recall: $E[g(X)] = \sum_x g(x) p_X(x)$

- **Second moment:** $E[X^2] = \sum_x x^2 p_X(x)$

- **Variance**

$$\text{var}(X) = E[(X - E[X])^2]$$

$$= \sum_x (x - E[X])^2 p_X(x)$$

$$= E[X^2] - (E[X])^2$$

Properties:

- $\text{var}(X) \geq 0$

- $\text{var}(\alpha X + \beta) = \alpha^2 \text{var}(X)$

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Recitation 5
September 23, 2010

1. (a) Derive the expected value rule for functions of random variables $\mathbf{E}[g(X)] = \sum_x g(x)p_X(x)$.
(b) Derive the property for the mean and variance of a linear function of a random variable $Y = aX + b$.

$$\mathbf{E}[Y] = a\mathbf{E}[X] + b, \quad \text{var}(Y) = a^2\text{var}(X).$$

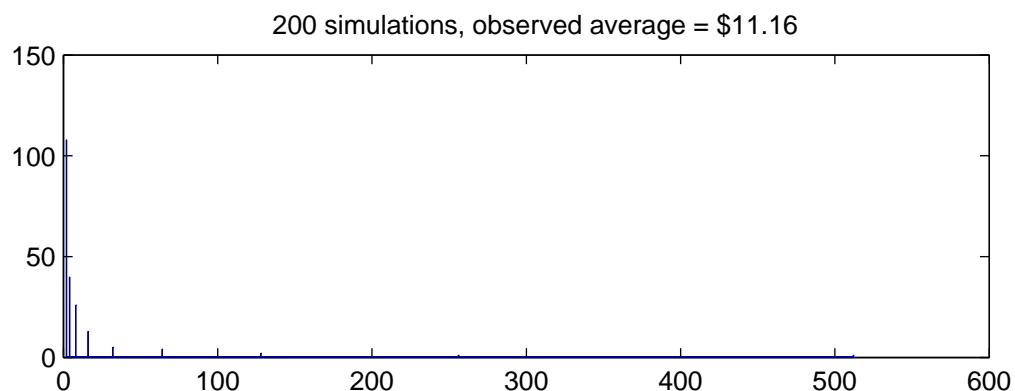
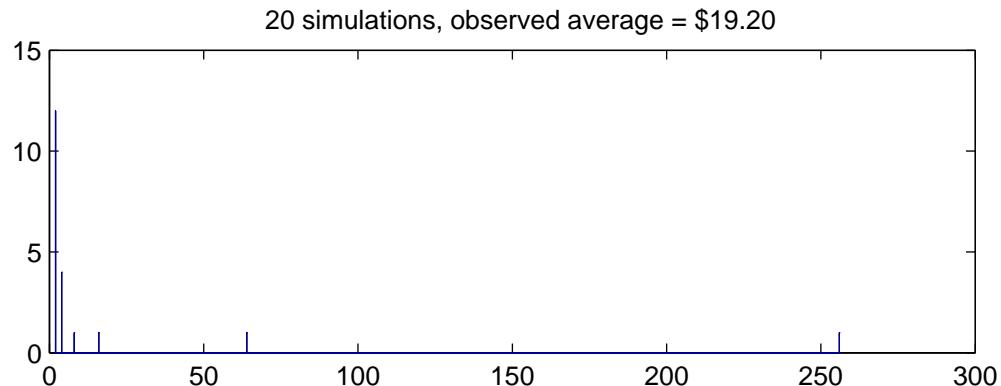
- (c) Derive $\text{var}(X) = \mathbf{E}[X^2] - (\mathbf{E}[X])^2$
2. A marksman takes 10 shots at a target and has probability 0.2 of hitting the target with each shot, independently of all other shots. Let X be the number of hits.

- (a) Calculate and sketch the PMF of X .
(b) What is the probability of scoring no hits?
(c) What is the probability of scoring more hits than misses?
(d) Find the expectation and the variance of X .
(e) Suppose the marksman has to pay \$3 to enter the shooting range and he gets \$2 dollars for each hit. Let Y be his profit. Find the expectation and the variance of Y .
(f) Now let's assume that the marksman enters the shooting range for free and gets the number of dollars that is equal to the square of the number of hits. Let Z be his profit. Find the expectation of Z .
3. 4 buses carrying 148 job-seeking MIT students arrive at a job convention. The buses carry 40, 33, 25, and 50 students, respectively. One of the students is randomly selected. Let X denote the number of students that were on the bus carrying this randomly selected student. One of the 4 bus drivers is also randomly selected. Let Y denote the number of students on his bus.
 - (a) Which of $E[X]$ or $E[Y]$ do you think is larger? Give your reasoning in words.
(b) Compute $E[X]$ and $E[Y]$.
4. Problem 2.21, page 123 in the text.

St. Petersburg paradox. You toss independently a fair coin and you count the number of tosses until the first tail appears. If this number is n , you receive 2^n dollars. What is the expected amount that you will receive? How much would you be willing to pay to play this game?

Recitation 5: Extra Handout
September 23, 2010

1. To show some relevant computations to Problem 4, the results (plotted as histograms) of simulations of this game have been plotted below for various numbers of simulations.



MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

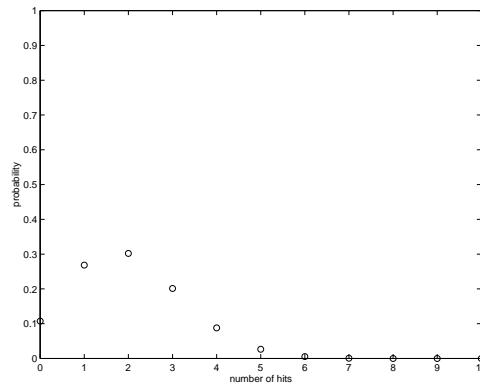
For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Recitation 5 Solutions
September 23, 2010

1. (a) See derivation in textbook pp. 84-85.
 (b) See derivation in textbook p. 86.
 (c) See derivation in textbook p. 87.
2. (a) X is a Binomial random variable with $n = 10$, $p = 0.2$. Therefore,

$$p_X(k) = \binom{10}{k} 0.2^k 0.8^{10-k}, \quad \text{for } k = 0, \dots, 10$$

and $p_X(k) = 0$ otherwise.



- (b) $\mathbf{P}(\text{No hits}) = p_X(0) = (0.8)^{10} = \boxed{0.1074}$
- (c) $\mathbf{P}(\text{More hits than misses}) = \sum_{k=6}^{10} p_X(k) = \sum_{k=6}^{10} \binom{10}{k} 0.2^k 0.8^{10-k} = \boxed{0.0064}$
- (d) Since X is a Binomial random variable,

$$\mathbf{E}[X] = 10 \cdot 0.2 = \boxed{2} \quad \text{var}(X) = 10 \cdot 0.2 \cdot 0.8 = \boxed{1.6}$$

- (e) $Y = 2X - 3$, and therefore

$$\mathbf{E}[Y] = 2\mathbf{E}[X] - 3 = \boxed{1} \quad \text{var}(Y) = 4\text{var}(X) = \boxed{6.4}$$

- (f) $Z = X^2$, and therefore

$$\mathbf{E}[Z] = \mathbf{E}[X^2] = (\mathbf{E}[X])^2 + \text{var}(X) = \boxed{5.6}$$

3. (a) We expect $\mathbf{E}[X]$ to be higher than $\mathbf{E}[Y]$ since if we choose the student, we are more likely to pick a bus with more students.
- (b) To solve this problem formally, we first compute the PMF of each random variable and then compute their expectations.

$$p_X(x) = \begin{cases} 40/148 & x = 40 \\ 33/148 & x = 33 \\ 25/148 & x = 25 \\ 50/148 & x = 50 \\ 0 & \text{otherwise.} \end{cases}$$

and $\mathbf{E}[X] = 40 \frac{40}{148} + 33 \frac{33}{148} + 25 \frac{25}{148} + 50 \frac{50}{148} = 39.28$

$$p_Y(y) = \begin{cases} 1/4 & y = 40, 33, 25, 50 \\ 0 & \text{otherwise.} \end{cases}$$

and $\mathbf{E}[Y] = 40 \frac{1}{4} + 33 \frac{1}{4} + 25 \frac{1}{4} + 50 \frac{1}{4} = 37$

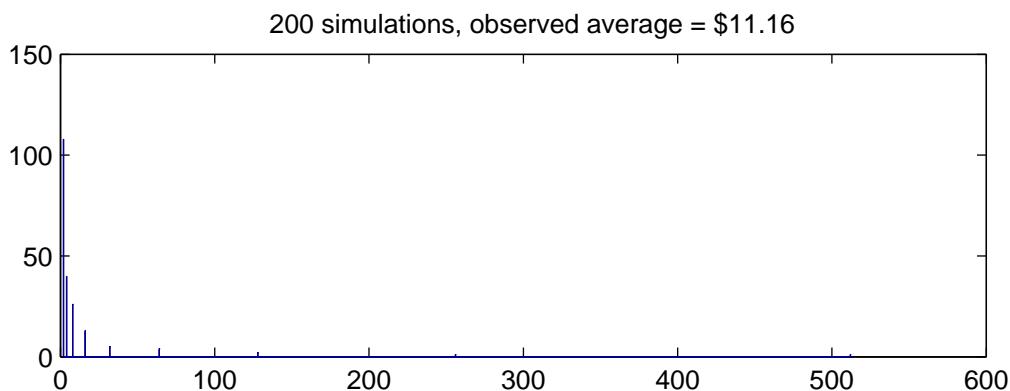
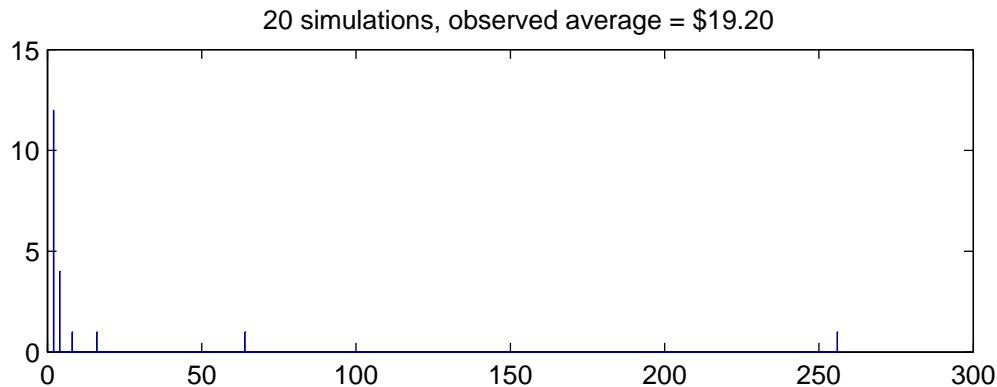
Clearly, $\mathbf{E}[X] > \mathbf{E}[Y]$.

4. The expected value of the gain for a single game is infinite since if X is your gain, then

$$\sum_{k=1}^{\infty} 2^k \cdot 2^{-k} = \sum_{k=1}^{\infty} 1 = \infty$$

Thus if you are faced with the choice of playing for given fee f or not playing at all, and your objective is to make the choice that maximizes your expected net gain, you would be willing to pay any value of f . However, this is in strong disagreement with the behavior of individuals. In fact experiments have shown that most people are willing to pay only about \$20 to \$30 to play the game. The discrepancy is due to a presumption that the amount one is willing to pay is determined by the expected gain. However, expected gain does not take into account a persons attitude towards risk taking.

Below are histograms showing the payout results for various numbers of simulations of this game:



MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Tutorial 2
September 23/24, 2010

1. A player is randomly dealt 13 cards from a standard 52-card deck.
 - (a) What is the probability the 13th card dealt is a king?
 - (b) What is the probability the 13th card dealt is the first king dealt?

2. Consider a random variable X such that

$$p_X(x) = \frac{x^2}{a} \text{ for } x \in \{-3, -2, -1, 1, 2, 3\}, \quad \mathbf{P}(X = x) = 0 \text{ for } x \notin \{-3, -2, -1, 1, 2, 3\},$$

where $a > 0$ is a real parameter.

- (a) Find a .
 - (b) What is the PMF of the random variable $Z = X^2$?
3. 90 students, including Joe and Jane, are to be split into three classes of equal size, and this is to be done at random. What is the probability that Joe and Jane end up in the same class?
 4. Draw the top 7 cards from a well-shuffled standard 52-card deck. Find the probability that the 7 cards include exactly 3 aces.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Tutorial 2 Solutions
September 23/24, 2010

1. A player is randomly dealt 13 cards from a standard 52-card deck.

- (a) What is the probability the 13th card dealt is a king?

Answer: $\frac{4}{52}$.

Solution: Since we are not told anything about the first 12 cards that are dealt, the probability that the 13th card dealt is a King, is the same as the probability that the first card dealt, or in fact any particular card dealt is a King, and this equals: $\frac{4}{52}$.

- (b) What is the probability the 13th card dealt is the first king dealt?

Answer: $\frac{1}{13} \cdot 4 \binom{48}{12} / \binom{52}{13}$.

Solution: The probability that the 13th card is the first king to be dealt is the probability that out of the first 13 cards to be dealt, exactly one was a king, and that the king was dealt last. Now, given that exactly one king was dealt in the first 13 cards, the probability that the king was dealt last is just $1/13$, since each “position” is equally likely. Thus, it remains to calculate the probability that there was exactly one king in the first 13 cards dealt. To calculate this probability we count the “favorable” outcomes and divide by the total number of possible outcomes. We first count the favorable outcomes, namely those with exactly one king in the first 13 cards dealt. We can choose a particular king in 4 ways, and we can choose the other 12 cards in $\binom{48}{12}$ ways, therefore there are $4 \cdot \binom{48}{12}$ favorable outcomes. There are $\binom{52}{13}$ total outcomes, so the desired probability is

$$\frac{1}{13} \cdot \frac{4 \binom{48}{12}}{\binom{52}{13}}.$$

For an alternative solution, we argue as in Example 1.10. The probability that the first card is not a king is $48/52$. Given that, the probability that the second is not a king is $47/51$. We continue similarly until the 12th card. The probability that the 12th card is not a king, given that none of the preceding 11 was a king, is $37/41$. (There are $52 - 11 = 41$ cards left, and $48 - 11 = 37$ of them are not kings.) Finally, the conditional probability that the 13th card is a king is $4/40$. The desired probability is

$$\frac{48 \cdot 47 \cdots 37 \cdot 4}{52 \cdot 51 \cdots 41 \cdot 40}.$$

2. Consider a random variable X such that

$$p_X(x) = \frac{x^2}{a} \text{ for } x \in \{-3, -2, -1, 1, 2, 3\}, \quad \mathbf{P}(X = x) = 0 \text{ for } x \notin \{-3, -2, -1, 1, 2, 3\},$$

where $a > 0$ is a real parameter.

- (a) Find a .

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
 (Fall 2010)

Solution. The sum of the values of the PMF of a random variable over all values that it takes with positive probability must be equal to 1. Hence, we have

$$\begin{aligned} 1 &= \sum_{x=-3}^3 p_X(x) \\ &= \frac{9}{a} + \frac{4}{a} + \frac{1}{a} + \frac{1}{a} + \frac{4}{a} + \frac{9}{a} \\ &= \frac{28}{a}, \end{aligned}$$

which implies that $a = 28$.

- (b) What is the PMF of the random variable $Z = X^2$?

Solution. The following table shows the value of Z for a given value of X and the probability of that event.

x	-3	-2	-1	1	2	3
$p_X(x)$	9/28	1/7	1/28	1/28	1/7	9/28
$Z X=x$	9	4	1	1	4	9

We see that Z can take only three possible values with non-zero probability, namely 1, 4, and 9. In addition, for each value, there correspond two values of X . So we have, for example, $p_Z(9) = \mathbf{P}(Z = 9) = \mathbf{P}(X = -3) + \mathbf{P}(X = 3) = p_X(-3) + p_X(3)$. Hence the PMF of Z is given by

$$p_Z(z) = \begin{cases} 1/14 & \text{if } z = 1, \\ 2/7 & \text{if } z = 4, \\ 9/14 & \text{if } z = 9. \end{cases}$$

3. Suppose we label the classes A , B , and C . Now the probability that Joe and Jane will both be in class A is the number of possible combinations for class A that involve both Joe and Jane, divided by the total number of combinations for class A . Therefore the probability we are after is:

$$\frac{\binom{88}{28}}{\binom{90}{30}}.$$

Since there are three classrooms, the probability that Joe and Jane end up in the same classroom is simply three times the answer we found above:

$$3 \cdot \frac{\binom{88}{28}}{\binom{90}{30}}.$$

Another way of looking at the problem is described as follows,

Assume one of them pick first, say Joe. He can pick any one of the 90 available places. Then it's Jane's turn to pick. She has a probability of $\frac{29}{89}$ of picking in the same class as Joe. Therefore, the overall probability is $\frac{29}{89}$, which is the same as $3 \cdot \frac{\binom{88}{28}}{\binom{90}{30}}$.

4. Let A = event the 7 cards include exactly 3 aces.

$$P(A) = \frac{(\# \text{ ways to choose 3 aces}) \cdot (\# \text{ ways to choose other 4 cards})}{\# \text{ ways to choose 7 cards}} = \frac{\binom{4}{3} \binom{48}{4}}{\binom{52}{7}}.$$

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Tutorial 3
September 30/October 1, 2010

1. Let X and Y be independent random variables. Random variable X has mean μ_X and variance σ_X^2 , and random variable Y has mean μ_Y and variance σ_Y^2 . Let $Z = 2X - 3Y$. Find the mean and variance of Z in terms of the means and variances of X and Y .
2. Problem 2.40, page 133 in the text.
A particular professor is known for his arbitrary grading policies. Each paper receives a grade from the set $\{A, A-, B+, B, B-, C+\}$, with equal probability, independently of other papers. How many papers do you expect to hand in before you receive each possible grade at least once?
3. The joint PMF of the random variables X and Y is given by the following table:

$y = 3$	c	c	$2c$
$y = 2$	$2c$	0	$4c$
$y = 1$	$3c$	c	$6c$
	$x = 1$	$x = 2$	$x = 3$

- (a) Find the value of the constant c .
- (b) Find $p_Y(2)$.
- (c) Consider the random variable $Z = YX^2$. Find $\mathbf{E}[Z \mid Y = 2]$.
- (d) Conditioned on the event that $X \neq 2$, are X and Y independent? Give a one-line justification.
- (e) Find the conditional variance of Y given that $X = 2$.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Tutorial 3: Solutions

1. In general we have that $\mathbf{E}[aX + bY + c] = a\mathbf{E}[X] + b\mathbf{E}[Y] + c$. Therefore,

$$\mathbf{E}[Z] = 2 \cdot \mathbf{E}[X] - 3 \cdot \mathbf{E}[Y].$$

For the case of independent random variables, we have that if $Z = a \cdot X + b \cdot Y$, then

$$\text{var}(Z) = a^2 \cdot \text{var}(X) + b^2 \cdot \text{var}(Y).$$

Therefore, $\text{var}(Z) = 4 \cdot \text{var}(X) + 9 \cdot \text{var}(Y)$.

2. See online solutions.
3. (a) We can find c knowing that the probability of the entire sample space must equal 1.

$$\begin{aligned} 1 &= \sum_{x=1}^3 \sum_{y=1}^3 p_{X,Y}(x,y) \\ &= c + c + 2c + 2c + 4c + 3c + c + 6c \\ &= 20c \end{aligned}$$

Therefore, $c = \frac{1}{20}$.

$$(b) p_Y(2) = \sum_{x=1}^3 p_{X,Y}(x,2) = 2c + 0 + 4c = 6c = \frac{3}{10}.$$

$$(c) Z = YX^2$$

$$\begin{aligned} \mathbf{E}[Z \mid Y = 2] &= \mathbf{E}[YX^2 \mid Y = 2] \\ &= \mathbf{E}[2X^2 \mid Y = 2] \\ &= 2\mathbf{E}[X^2 \mid Y = 2] \end{aligned}$$

$$p_{X|Y}(x \mid 2) = \frac{p_{X,Y}(x,2)}{p_Y(2)}.$$

Therefore,

$$p_{X|Y}(x \mid 2) = \begin{cases} \frac{1/10}{3/10} = \frac{1}{3} & \text{if } x = 1 \\ \frac{1/5}{3/10} = \frac{2}{3} & \text{if } x = 3 \\ 0 & \text{otherwise} \end{cases}$$

$$\begin{aligned} \mathbf{E}[Z \mid Y = 2] &= 2 \sum_{x=1}^3 x^2 p_{X|Y}(x \mid 2) \\ &= 2 \left((1^2) \cdot \frac{1}{3} + (3^2) \cdot \frac{2}{3} \right) \\ &= \frac{38}{3} \end{aligned}$$

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
 (Fall 2010)

(d) Yes. Given $X \neq 2$, the distribution of X is the same given $Y = y$.

$$\mathbf{P}(X = x | Y = y, X \neq 2) = \mathbf{P}(X = x | X \neq 2).$$

For example,

$$\mathbf{P}(X = 1 | Y = 1, X \neq 2) = \mathbf{P}(X = 1 | Y = 3, X \neq 2) = \mathbf{P}(X = 1 | X \neq 2) = \frac{1}{3}$$

(e) $p_{Y|X}(y | 2) = \frac{p_{X,Y}(2,y)}{p_X(2)}$.

$$p_X(2) = \sum_{y=1}^3 p_{X,Y}(2,y) = c + 0 + c = 2c = \frac{1}{10}.$$

Therefore,

$$p_{Y|X}(y | 2) = \begin{cases} \frac{1/20}{1/10} = \frac{1}{2} & \text{if } y = 1 \\ \frac{1/20}{1/10} = \frac{1}{2} & \text{if } y = 3 \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbf{E}[Y^2 | X = 2] = \sum_{y=1}^3 y^2 p_{Y|X}(y | 2) = (1^2) \cdot \frac{1}{2} + (3^2) \cdot \frac{1}{2} = 5.$$

$$\mathbf{E}[Y | X = 2] = \sum_{y=1}^3 y p_{Y|X}(y | 2) = 1 \cdot \frac{1}{2} + 3 \cdot \frac{1}{2} = 2.$$

$$\text{var}(Y | X = 2) = \mathbf{E}[Y^2 | X = 2] - \mathbf{E}[Y | X = 2]^2 = 5 - 2^2 = 1.$$

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2010

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013

Transcript – Recitation: Flipping a Coin a Random Number of Times

In this problem, we're looking at a two stage process in which the first stage, we roll a fair die which has four faces to obtain a number N , where N belongs to the set 0, 1, 2, and 3 with equal probability. Now, given the result of the die roll, N will toss a fair coin N times in getting K heads from the coin tosses.

For instance, if from the first die roll, we get N equal to 3, then we'll toss a coin 3 times. Let's say the outcome is heads, heads, and tails. And that will give us K equal to 2.

For part A, we're asked to compute the PMF for N , which is a result of the first die roll. Now, since we had assumed the die roll was uniformly distributed in the set in the set 0, 1, 2, and 3, we have that the chance of N being equal to any little n is equal to $1/4$ if n is in the set 0, 1, 2, 3, and 0 otherwise. If we were to plot this in a figure, we'll have the following plot.

For part B, things are getting a little more complicated. This time, we want to compute the joint PMF between N and K for N equal to little n and K equal to little k . What we'll do first is to use the law of conditional probability to break the joint probability into the product of probability of K is equal to little k conditional on N is equal to little n , multiply by the probability that N is equal to little n .

Now, the second term right here is simply the PMF of N , which will be computed earlier. So this gives us $1/4$ times probability K equal to little k , N equal to little n , for all N in the set 0, 1, 2, and 3. Now, clearly if N is not one of those four values, this whole event couldn't have happened in the first place, and hence will have P and K equal to 0.

We'll now go over all the cases for little n in this expression right here. The first case is the simplest. If we assume that little n is equal to 0, that means the die roll was 0, and hence we're not tossing any coins afterwards. And this implies that we must have K is equal to 0, which, mathematically speaking, is equivalent to saying probability of K equal to 0 conditional on N equal to 0 is 1. And K being any other value conditional N equal to 0 is 0. So we're done with the case that little n is equal to 0.

Now, let's say little n is in the set 1, 2, and 3. In this case, we want to notice that after having observed the value of N , all the coin tosses for N times are conditionally independent from each other. What this means is now the total number of heads in the subsequent coin toss is equal in distribution to a binomial random variable with parameter n and $1/2$. And here says the number of trials is n , and $1/2$ is because the coin is fair. And the reason it is a binomial random variable, again, is because the coin tosses are independent conditional on the outcome of the die roll.

And now we're done, since we know what the binomial distribution looks like given parameter n and $1/2$. And we'll simply substitute based on the case of n the conditional distribution of K back into the product we had earlier, which in turn will give us the joint PMF.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

This table summarizes the PMF we were computing earlier. P of N , K , little n , and little k . Now, as we saw before, if n equal to 0, the only possibility for k is equal to 0. And this is scaled by the probability of n equal to 0, which is $1/4$. For any other values of n , we see that the distribution of k , conditional n , is the same as a binomial random variable with n trials. And again, every entry here is scaled by $1/4$. And this completes part B.

In part C, we're asked for the conditional PMF of K conditioning on the value of N being equal to 2. Now, as we discussed in part B, when N is equal to 2, we're essentially flipping a fair coin twice, and this should give us the same distribution as a binomial random variable with parameter 2 and $1/2$. Now, 2 is the number of flips, and $1/2$ is the chance of seeing a head in each flip. And that gives us the following distribution.

But there's another way to see this. It's to write $P K$ given N , little k , and you go to 2 by using the law of conditional probability as $P K, N$, the joint PMF, $k n2$, divided by the probability that N is equal to 2.

Now, we know that probability n equal to 2 is simply $1/4$, so this gives us 4 times the joint density $K, N, k, 2$. In other words, in order to arrive at the distribution right here, [INAUDIBLE] to go back to the table we had earlier and look at the role where n is equal to 2 and multiply each number by 4.

Finally, in part D, we're asked for the conditional distribution of N , write as $P N$, given K of N equal to little n conditional on K is equal to 2. Again, we'll apply the formula for conditional probability. This is equal to the joint PMF evaluated at n and 2 divided by the probability of K being equal to 2.

Since we have computed the entire table of the joint PMF, this shouldn't be too difficult. In particular, for the denominator, the probability that k is ever equal to 2, we just look at the column right here. So the entries in this column shows all the cases where k can be equal to 2. And in fact, we can see that k can be equal to 2 only if n is equal to 2 or 3. Clearly, if you toss the coin fewer than 2 times, there's no chance that we'll get 2 heads. So to get this probability right here, we'll add up the number in these two cells. So we get $P N, K$, little n , and 2 divided by $1/16$ plus $3/32$.

Now, the numerator, again, can be read off from the table right here. In particular, this tells us that there are only two possibilities. Either n is equal to 2 or n equal to 3. When n is equal to 2, we know this quantity gives us $1/16$ reading off this cell divided by $1/16$ plus $3/32$ for n equal to 2. And the remaining probability goes to the case where n is equal to 3.

So this is 3 divided by 32 , $1/16$ plus $3/32$, which simplifies to $2/5$ and $3/5$. And this distribution gives us the following plot. And this completes our problem.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013 Transcript – Recitation: Joint Probability Mass Function (PMF) Drill 1

Welcome back guys. Today we're going to work on a problem that tests your knowledge of joint PMFs. And we're also going to get some practice computing conditional expectations and conditional variances. So in this problem, we are given a set of points in the xy plane. And we're told that these points are equally likely.

So there's eight of them. And each point has a probability of $1/8$ of occurring. And we're also given this list of questions. And we're going to work through them together.

So in part a, we are asked to find the values of x that maximize the conditional expectation of y given x . So jumping right in, this is the quantity we're interested in. And so this quantity is a function of x .

You plug-in various values of x . And then this will spit out a scalar value. And that value will correspond to the conditional expectation of y conditioned on the value of x that you put in. So let's see, when x is equal to 0, for instance, let's figure out what this value is.

Well, when x is equal to 0 we're living in a world, essentially, on this line. So that means that only these two points could have occurred. And in particular, y can only take on the values of 1 and 3.

Now, since all these points in the unconditional universe were equally likely, in the conditional universe they will still be equally likely. So this happens with probability $1/2$. And this happens with probability $1/2$.

And therefore, the expectation would just be $3/2$ plus $1/2$ which is $4/2$, or 2. But a much faster way of seeing this-- and it's the strategy that I'm going to use for the rest of the problem-- is to remember that expectation acts like center of mass. So the center of mass, when these two points are equally likely, is just the midpoint, which of course is 2.

So we're going to use that intuition on the other ones. So I'm skipping to x is equal to 2 because 1 and 3 are not possible. So when x is equal to 2, y can only take on the values of 1 or 2. Again, they're equally likely. So the center of mass is in the middle which happens at 1.5 or $3/2$.

Similarly, x is equal to 4. We're living in this conditional universe, where y can take on of these four points with probability $1/4$ each. And so again, we expect the center of mass to be at 1.5 or $3/2$. And this quantity is undefined otherwise.

OK, so we're almost done. Now we just need to find which value of x maximizes this. Well, let's see, 2 is the biggest quantity out of all of these numbers. So the maximum is 2. And it occurs when x is equal to 0.

So we come over here. And we found our answer. x is equal to 0 is the value, which maximizes the conditional expectation of y given x .

So part b is very similar to part a. But there is slightly more computation involved. Because now we're dealing with the variance and not an expectation. And variance is usually a little bit tougher to compute.

So we're going to start in the same manner. But I want you guys to see if you can figure out intuitively what the right value is. I'm going to do the entire computation now. And then you can compare whether your intuition matches with the real results.

So variance of x conditioned on a particular value of y , this is now a function of y . For each value of y you plug in you're going to get out a scalar number. And that number represents the conditional variance of x when you condition on the value of y that you plugged in.

So let's see, when y is equal to 0 we have a nice case. If y is equal to 0 we have no freedom about what x is. This is the only point that could have occurred.

Therefore, x definitely takes on a value of 4. And there's no uncertainty left. So in other words, the variance is 0.

Now, if y is equal to 1, x can take on a value of 0, a value of 2 or a value of 4. And these all have the same probability of occurring, of 1/3,

And again, the reasoning behind that is that all eight points were equally likely in the unconditional universe. If you condition on y being equal to 1 these outcomes still have the same relative frequency. Namely, they're still equally likely.

And since there are three of them they now have a probability of 1/3 each. So we're going to go ahead and use a formula that hopefully, you guys remember. So in particular, variance is the expectation of x squared minus the expectation of x all squared, the whole thing squared.

So let's start by computing this number first. So conditioned on y is equal to 1-- so we're in this line-- the expectation of x is just 2, right? The same center-of-mass to argument.

So this, we have a minus 2 squared over here. Now, x squared is only slightly more difficult. With probability 1/3, x squared will take on a value of 0.

With probability 1/3, x squared will take on a value of 4. I'm just doing 2 squared. And with probability 1/3, x squared takes on a value of 4 squared or 16.

So writing down when I just said, we have 0 times 1/3 which is 0. We have 2 squared, which is 4 times 1/3. And then we have 4 squared, which is 16 times 1/3.

And then we have our minus 4 from before. So doing this math out, we get, let's see, 20/3 minus 12/3, which is equal to 8/3, or 8/3. So we'll come back up here and put 8/3.

So I realize I'm going through this pretty quickly. Hopefully this step didn't confuse you. Essentially, what I was doing is, if you think of x squared as a new random variable, x squared, the possible values that it can take on are 0, 4, and 16 when you're conditioning on y is equal to 1. And so I was simply saying that that random variable takes on those values with equal probability. So let's move on to the next one.

So if we condition on y is equal to 2 we're going to do a very similar computation. Oops, I shouldn't have erased that. OK, so we're going to use the same formula that we just used, which is the expectation of x given y is equal to 2. Sorry, x squared minus the expectation of x conditioned on y is equal to 2, all squared.

So conditioned on y is equal to 2, the expectation of x is 3. Same center of mass argument. So 3 squared is 9.

And then x squared can take on a value of 4. Or it can take on a value of 16. And it does so with equal probability.

So we get $4/2$, 4 plus 16 over 2. So this is 2 plus 8, which is 10, minus 9. That'll give us 1.

So we get a 1 when y is equal to 2. And last computation and then we're done. I'm still recycling the same formula. But now we're conditioning on y is equal to 3. And then we'll be done with this problem, I promise.

OK, so when y is equal to 3 x can take on the value of 0. Or it can take on the value of 4. Those two points happen with probability $1/2$, $1/2$. So the expectation is right in the middle which is 2. So we get a minus 4.

And similarly, x squared can take on the value of 0. When x takes on the value of 0-- and that happens with probability $1/2$ -- similarly, x squared can take on the value of 16 when x takes on the value of 4. And that happens with probability $1/2$.

So we just have $0/2$ plus $16/2$ minus 4. And this gives us 8 minus 4, which is simply 4. So finally, after all that computation, we are done. We have the conditional variance of x given y .

Again, we're interested in when this value is largest. And we see that 4 is the biggest value in this column. And this value occurs when y takes on a value of 3. So our answer, over here, is y is equal to 3.

All right, so now we're going to switch gears in part c and d a little bit. And we're going to be more concerned with PMFs, et cetera. So in part c, we're given a random variable called r which is defined as the minimum of x and y .

So for instance, this is the 0.01. The minimum of 0 and 1 is 0. So r would have a value of 0 here. Now, we can be a little bit smarter about this.

If we plot the line, y is equal to x . So that looks something like this. We see that all of the points below this line satisfy y being less or equal to x . And all the points above this line have y greater than or equal to x .

So if y is less than or equal to x , you hopefully agree that here the min, or r , is equal to y . But over here, the min, r , is actually equal to x , since x is always smaller.

So now we can go ahead quickly. And I'm going to write the value of r next each point using this rule. So here, r is the value of y , which is 1.

Here, r is equal to 0. Here r is 1. Here r is 2. Here r is 3.

Over here, r is the value of x . So r is equal to 0. And r is equal to 0 here.

And so the only point we didn't handle is the one that lies on the line. But in that case it's easy. Because x is equal to 2. And y is equal to 2. So the min is simply 2. So with this information I claim we're now done. We can just write down what the PMF of r is.

So in particular, r takes on a value of 0. When this point happens, this point happens, or this point happens. And those collectively have a probability of $3/8$ of occurring.

r can take on a value of 1 when either of these two points happen. So that happens with probability $2/8$. r is equal to 2. This can happen in two ways. So we get $2/8$. And r equal to 3 can happen in only one way. So we get $1/8$.

Quick sanity check, 3 plus 2 is 5, plus 2 is 7, plus 1 is 8. So our PMF sums to 1. And to be complete, we should sketch it. Because the problem asks us to sketch it.

So we're plotting PR of r , 0, 1, 2, 3. So here we get, let's see, 1, 2, 3. For 0 we have $3/8$. For 1 we have $2/8$. For 2 we have $2/8$. And for 3 we have $1/8$.

So this is our fully labeled sketch of \Pr of r . And forgive me for erasing so quickly, but you guys can pause the video, presumably, if you need more time. Let's move on to part d.

So in part d we're given an event named a , which is the event that x squared is greater than or equal to y . And then we're asked to find the expectation of xy in the unconditional universe. And then the expectation of x times y conditioned on a .

So let's not worry about the conditioning for now. Let's just focus on the unconditional expectation of x times y . So I'm just going to erase all these r 's so I don't get confused.

But we're going to follow a very similar strategy, which is at each point I'm going to label what the value of w is. And we'll find the expectation of w that way.

So let's see, here, we have 4 times 0. So w is equal to 0. Here we have 4 times 1.

w is equal to 4. 4 times 2, w is equal to 8. 4 times 3, w is equal to 12.

w is equal to 2. w is equal to 4. w is equal to 0. w is equal to 0.

OK, so that was just algebra. And now, I claim again, we can just write down what the expectation of x times y is. And I'm sorry, I didn't announce my notation. I should mention that now.

I was defining w to be the random variable x times y. And that's why I labeled the product of x times y as w over here. My apologies about not defining that random variable. So the expectation of w, well, w takes on a value of 0. When this happens, this happens or that happens. And we know that those three points occur with probability 3/8.

So we have 0 times 3/8. I'm just using the normal formula for expectation. w takes on a value of 2 with probability 1/8. Because this is the lead point in which it happens, 2 times 1/8.

Plus it can take on the value of 4 with probability 2/8, 4 times 2/8. And 8, with 1/8 probability. And similarly, 12 with 1/8 probability.

So this is just algebra. The numerator sums up to 30. Yes, that's correct. So we have 30/8, which is equal to 15/4. So this is our first answer for part d.

And now we have to do this slightly trickier one, which is the conditional expectation of x times y, or w conditioned on a. So similar to what I did in part c, I'm going to draw the line y equals x squared. So y equals x squared is 0 here, 1 here. And at 2, it should take on a value of 4.

So the curve should look something like this. This is the line y is equal to x squared. So we know all the points below this line satisfy y less than or equal to x squared. And all the points above this line have y greater than or equal to x squared.

And a is y less than or equal to x squared. So we are in the conditional universe where only points below this line can happen. So that one, that one, that one, that one, that one and that one. So there are six of them.

And again, in the unconditional world, all of the points were equally likely. So in the conditional world these six points are still equally likely. So they each happen with probability 1/6.

So in this case, the expectation of w is simply 2 times 1/6 plus 0 times 1/6. But that's 0. So I'm not going to write it.

4 times 2/6 plus 4 times 2/6 plus 8 times 1/6, plus 12 times 1 over 6. And again, the numerator summed to 30. But this time our denominator is 6. So this is simply 5.

So we have, actually, finished the problem. Because we've computed this value and this value. And so the important takeaways of this problem are, essentially, honestly, just to get you comfortable with computing things involving joint PMFs.

We talked a lot about finding expectations quickly by thinking about center of mass and the geometry of the problem. We've got practice computing conditional variances. And we did some derived distributions. And we'll do a lot more of those later.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013

Transcript – Lecture 6

The following content is provided under a Creative Commons license. Your support will help MIT OpenCourseWare continue to offer high-quality educational resources for free. To make a donation or view additional materials from hundreds of MIT courses, visit MIT OpenCourseWare at ocw.mit.edu.

PROFESSOR: OK so let's start. So today, we're going to continue the subject from last time. And the subject is random variables. As we discussed, random variables basically associate numerical values with the outcomes of an experiment. And we want to learn how to manipulate them.

Now to a large extent, what's going to happen, what's happening during this chapter, is that we are revisiting the same concepts we have seen in chapter one. But we're going to introduce a lot of new notation, but really dealing with the same kind of stuff. The only difference where we go beyond the new notation, the new concept in this chapter is the concept of the expectation or expected values. And we're going to learn how to manipulate expectations.

So let us start with a quick review of what we discussed last time. We talked about random variables. Loosely speaking, random variables are random quantities that result from an experiment.

More precisely speaking, mathematically speaking, a random variable is a function from the sample space to the real numbers. That is, you give me an outcome, and based on that outcome, I can tell you the value of the random variable. So the value of the random variable is a function of the outcome that we have.

Now given a random variable, some of the numerical outcomes are more likely than others. And we want to say which ones are more likely and how likely they are. And the way we do that is by writing down the probabilities of the different possible numerical outcomes.

Notice here, the notation. We use uppercase to denote the random variable. We use lowercase to denote real numbers.

So the way you read this, this is the probability that the random variable, capital X, happens to take the numerical value, little x. This is a concept that's familiar from chapter one. And this is just the new notation we will be using for that concept.

It's the Probability Mass Function of the random variable, capital X. So the subscript just indicates which random variable we're talking about. And it's the probability assigned to a particular outcome.

And we want to assign such probabilities for all possibly numerical values. So you can think of this as being a function of little x. And it tells you how likely every little x is going to be.

Now the new concept we introduced last time is the concept of the expected value for random variable, which is defined this way. You look at all the possible outcomes. And you form some kind of average of all the possible numerical values over the random variable capital X. You consider all the possible numerical values, and you form an average. In fact, it's a weighted average where, to every little x, you assign a weight equal to the probability that that particular little x is going to be realized.

Now, as we discussed last time, if you have a random variable, you can take a function of a random variable. And that's going to be a new random variable. So if capital X is a random variable and g is a function, g of X is a new random variable.

You do the experiment. You get an outcome. This determines the value of X. And that determines the value of g of X.

So the numerical value of g of X is determined by whatever happens in the experiment. It's random. And that makes it a random variable.

Since it's a random variable, it has an expectation of its own. So how would we calculate the expectation of g of X? You could proceed by just using the definition, which would require you to find the PMF of the random variable g of X. So find the PMF of g of X, and then apply the formula for the expected value of a random variable with known PMF.

But there is also a shortcut, which is just a different way of doing the counting and the calculations, in which we do not need to find the PMF of g of X. We just work with the PMF of the original random variable. And what this is saying is that the average value of g of X is obtained as follows.

You look at all the possible results, the X's, how likely they are. And when that particular X happens, this is how much you get. And so this way, you add these things up. And you get the average amount that you're going to get, the average value of g of X, where you average over the likelihoods of the different X's.

Now expected values have some properties that are always true and some properties that sometimes are not true. So the property that is not always true is that this would be the same as g of the expected value of X. So in general, this is not true. You cannot interchange function and expectation, which means you cannot reason on the average, in general.

But there are some exceptions. When g is a linear function, then the expected value for a linear function is the same as that same linear function of the expectation. So for linear functions, so for random variable, the expectation behaves nicely.

So this is basically telling you that, if X is degrees in Celsius, alpha X plus b is degrees in Fahrenheit, you can first do the conversion to Fahrenheit and take the average. Or you can find the average temperature in Celsius, and then do the conversion to Fahrenheit. Either is valid.

So the expected value tells us something about where is the center of the distribution, more specifically, the center of mass or the center of gravity of the PMF, when you plot it as a bar graph. Besides the average value, you may be interested in knowing how far will you be from the average, typically. So let's look at this quantity, X minus expected value of X .

This is the distance from the average value. So for a random outcome of the experiment, this quantity in here measures how far away from the mean you happen to be. This quantity inside the brackets is a random variable.

Why? Because capital X is random. And what we have here is capital X , which is random, minus a number. Remember, expected values are numbers.

Now a random variable minus a number is a new random variable. It has an expectation of its own. We can use the linearity rule, expected value of something minus something else is just the difference of their expected value. So it's going to be expected value of X minus the expected value over this thing.

Now this thing is a number. And the expected value of a number is just the number itself. So we get from here that this is expected value minus expected value. And we get zero.

What is this telling us? That, on the average, the assigned difference from the mean is equal to zero. That is, the mean is here. Sometimes X will fall to the right. Sometimes X will fall to the left. On the average, the average distance from the mean is going to be zero, because sometimes the realized distance will be positive, sometimes it will be negative. Positives and negatives cancel out.

So if we want to capture the idea of how far are we from the mean, just looking at the assigned distance from the mean is not going to give us any useful information. So if we want to say something about how far we are, typically, we should do something different. One possibility might be to take the absolute values of the differences. And that's a quantity that sometimes people are interested in. But it turns out that a more useful quantity happens to be the variance of a random variable, which actually measures the average squared distance from the mean.

So you have a random outcome, random results, random numerical value of the random variable. It is a certain distance away from the mean. That certain distance is random.

We take the square of that. This is the squared distance from the mean, which is again random. Since it's random, it has an expected value of its own. And that expected value, we call it the variance of X . And so we have this particular definition.

Using the rule that we have up here for how to calculate expectations of functions of a random variable, why does that apply? Well, what we have inside the brackets here is a function of the random variable, capital X . So we can apply this rule where g is this particular function. And we can use that to calculate the variance, starting with the PMF of the random variable X . And then we have a useful formula that's a nice shortcut, sometimes, if you want to do the calculation.

Now one thing that's slightly wrong with the variance is that the units are not right, if you want to talk about the spread of a distribution. Suppose that X is a random variable measured in meters. The variance will have the units of meters squared. So it's a kind of a different thing.

If you want to talk about the spread of the distribution using the same units as you have for X , it's convenient to take the square root of the variance. And that's something that we define. And we call it the standard deviation of X , or the standard deviation of the distribution of X . So it tells you the amount of spread in your distribution. And it is in the same units as the random variable itself that you are dealing with.

And we can just illustrate those quantities with an example that's about as simple as it can be. So consider the following experiment. You're going to go from here to New York, let's say, 200 miles.

And you have two alternatives. Either you'll get your private plane and go at a speed of 200 miles per hour, constant speed during your trip, or otherwise, you'll decide to walk really, really slowly, at the leisurely pace of one mile per hour. So you pick the speed at random by doing this experiment, by flipping a coin.

And with probability one-half, you do one thing. With probability one-half, you do the other thing. So your V is a random variable.

In case you're interested in how much time it's going to take you to get there, well, time is equal to distance divided by speed. So that's the formula. The time itself is a random variable, because it's a function of V , which is random. How much time it's going to take you depends on the coin flip that you do in the beginning to decide what speed you are going to have.

OK, just as a warm up, the trivial calculations. To find the expected value of V , you argue as follows. With probability one-half, V is going to be one. And with probability one-half, V is going to be 200. And so the expected value of your speed is 100.5.

If you wish to calculate the variance of V , then you argue as follows. With probability one-half, I'm going to travel at the speed of one, whereas, the mean is 100.5. So this is the distance from the mean, if I decide to travel at the speed of one.

We take that distance from the mean squared. That's one contribution to the variance. And with probability one-half, you're going to travel at the speed of 200, which is this much away from the mean. You take the square of that.

OK, so approximately how big is this number? Well, this is roughly 100 squared. That's also 100 squared. So approximately, the variance of this random variable is 100 squared.

Now if I tell you that the variance of this distribution is 10,000, it doesn't really help you to relate it to this diagram. Whereas, the standard deviation, where you take the square root, is more interesting. It's the square root of 100 squared, which is a 100.

And the standard deviation, indeed, gives us a sense of how spread out this distribution is from the mean. So the standard deviation basically gives us some indication about this spacing that we have here. It tells us the amount of spread in our distribution.

OK, now let's look at what happens to time. V is a random variable. T is a random variable. So now let's look at the expected values and all of that for the time.

OK, so the time is a function of a random variable. We can find the expected time by looking at all possible outcomes of the experiment, the V 's, weigh them according to their probabilities, and for each particular V , keep track of how much time it took us. So if V is one, which happens with probability one-half, the time it takes is going to be 200. If we travel at speed of one, it takes us 200 time units.

And otherwise, if our speed is equal to 200, the time is one. So the expected value of T is once more the same as before. It's 100.5.

So the expected speed is 100.5. The expected time is also 100.5. So the product of these expectations is something like 10,000.

How about the expected value of the product of T and V ? Well, T times V is 200. No matter what outcome you have in the experiment, in that particular outcome, T times V is total distance traveled, which is exactly 200. And so what do we get in this simple example is that the expected value of the product of these two random variables is different than the product of their expected values.

This is one more instance of where we cannot reason on the average. So on the average, over a large number of trips, your average time would be 100. On the average, over a large number of trips, your average speed would be 100.

But your average distance traveled is not 100 times 100. It's something else. So you cannot reason on the average, whenever you're dealing with non-linear things. And the non-linear thing here is that you have a function which is a product of stuff, as opposed to just linear sums of stuff.

Another way to look at what's happening here is the expected value of the time. Time, by definition, is 200 over the speed. Expected value of the time, we found it to be about a 100. And so expected value of 200 over V is about a 100.

But it's different from this quantity here, which is roughly equal to 2, and so 200. Expected value of V is about 100. So this quantity is about equal to two. Whereas, this quantity up here is about 100.

So what do we have here? We have a non-linear function of V . And we find that the expected value of this function is not the same thing as the function of the expected value. So again, that's an instance where you cannot interchange expected values and functions. And that's because things are non-linear.

OK, so now let us introduce a new concept. Or maybe it's not quite a new concept. So we discussed, in chapter one, that we have probabilities. We also have conditional probabilities.

What's the difference between them? Essentially, none. Probabilities are just an assignment of probability values to give different outcomes, given a particular model.

Somebody comes and gives you new information. So you come up with a new model. And you have a new probabilities. We call these conditional probabilities, but they taste and behave exactly the same as ordinary probabilities.

So since we can have conditional probabilities, why not have conditional PMFs as well, since PMFs deal with probabilities anyway. So we have a random variable, capital X. It has a PMF of its own.

For example, it could be the PMF in this picture, which is a uniform PMF that takes for possible different values. And we also have an event. And somebody comes and tells us that this event has occurred.

The PMF tells you the probability that capital X equals to some little x. Somebody tells you that a certain event has occurred that's going to make you change the probabilities that you assign to the different values. You are going to use conditional probabilities.

So this part, it's clear what it means from chapter one. And this part is just the new notation we're using in this chapter to talk about conditional probabilities. So this is just a definition.

So the conditional PMF is an ordinary PMF. But it's the PMF that applies to a new model in which we have been given some information about the outcome of the experiment. So to make it concrete, consider this event here.

Take the event that capital X is bigger than or equal to two. In the picture, what is the event A? The event A consists of these three outcomes.

OK, what is the conditional PMF, given that we are told that event A has occurred? Given that the event A has occurred, it basically tells us that this outcome has not occurred. There's only three possible outcomes now.

In the new universe, in the new model where we condition on A, there's only three possible outcomes. Those three possible outcomes were equally likely when we started. So in the conditional universe, they will remain equally likely.

Remember, whenever you condition, the relative likelihoods remain the same. They keep the same proportions. They just need to be re-scaled, so that they add up to one.

So each one of these will have the same probability. Now in the new world, probabilities need to add up to 1. So each one of them is going to get a probability of 1/3 in the conditional universe.

So this is our conditional model. So our PMF is equal to 1/3 for X equals to 2, 3 and 4. All right.

Now whenever you have a probabilistic model involving a random variable and you have a PMF for that random variable, you can talk about the expected value of that random variable. We defined expected values just a few minutes ago. Here, we're dealing with a conditional model and conditional probabilities.

And so we can also talk about the expected value of the random variable X in this new universe, in this new conditional model that we're dealing with. And this leads us to the definition of the notion of a conditional expectation. The conditional expectation is nothing but an ordinary expectation, except that you don't use the original PMF.

You use the conditional PMF. You use the conditional probabilities. It's just an ordinary expectation, but applied to the new model that we have to the conditional universe where we are told that the certain event has occurred.

So we can now calculate the condition expectation, which, in this particular example, would be 1/3. That's the probability of a 2, plus 1/3 which is the probability of a 3 plus 1/3, the probability of a 4. And then you can use your calculator to find the answer, or you can just argue by symmetry. The expected value has to be the center of gravity of the PMF we're working with, which is equal to 3.

So conditional expectations are no different from ordinary expectations. They're just ordinary expectations applied to a new type of situation or a new type of model. Anything we might know about expectations will remain valid about conditional expectations.

So for example, the conditional expectation of a linear function of a random variable is going to be the linear function of the conditional expectations. Or you can take any formula that you might know, such as the formula that expected value of X is equal to the-- sorry-- expected value of g of X is the sum over all X's of g of X times the PMF of X. So this is the formula that we already know about how to calculate expectations of a function of a random variable.

If we move to the conditional universe, what changes? In the conditional universe, we're talking about the conditional expectation, given that event A has occurred. And we use the conditional probabilities, given that A has occurred.

So any formula has a conditional counterpart. In the conditional counterparts, expectations get replaced by conditional expectations. And probabilities get replaced by conditional probabilities. So once you know the first formula and you know the general idea, there's absolutely no reason for you to memorize a formula like this one. You shouldn't even have to write it on your cheat sheet for the exam, OK?

OK, all right, so now let's look at an example of a random variable that we've seen before, the geometric random variable, and this time do something a little more interesting with it. Do you remember from last time what the geometric random variable is? We do coin flips. Each time

there's a probability of P of obtaining heads. And we're interested in the number of tosses we're going to need until we observe heads for the first time.

The probability that the random variable takes the value K , this is the probability that the first K appeared at the K -th toss. So this is the probability of K minus 1 consecutive tails followed by a head. So this is the probability of having to weight K tosses.

And when we plot this PMF, it has this kind of shape, which is the shape of a geometric progression. It starts at 1, and it goes all the way to infinity. So this is a discrete random variable that takes values over an infinite set, the set of the positive integers.

So it's a random variable, therefore, it has an expectation. And the expected value is, by definition, we'll consider all possible values of the random variable. And we weigh them according to their probabilities, which leads us to this expression.

You may have evaluated that expression some time in your previous life. And there are tricks for how to evaluate this and get a closed-form answer. But it's sort of an algebraic trick. You might not remember it. How do we go about doing this summation?

Well, we're going to use a probabilistic trick and manage to evaluate the expectation of X , essentially, without doing any algebra. And in the process of doing so, we're going to get some intuition about what happens in coin tosses and with geometric random variables. So we have two people who are going to do the same experiment, flip a coin until they obtain heads for the first time.

One of these people is going to use the letter Y to count how many heads it took. So that person starts flipping right now. This is the current time. And they are going to obtain tails, tails, tails, until eventually they obtain heads. And this random variable Y is, of course, geometric, so it has a PMF of this form.

OK, now there is a second person who is doing that same experiment. That second person is going to take, again, a random number, X , until they obtain heads for the first time. And of course, X is going to have the same PMF as Y .

But that person was impatient. And they actually started flipping earlier, before the Y person started flipping. They flipped the coin twice. And they were unlucky, and they obtained tails both times. And so they have to continue.

Looking at the situation at this time, how do these two people compare? Who do you think is going to obtain heads first? Is one more likely than the other?

So if you play at the casino a lot, you'll say, oh, there were two tails in a row, so a head should be coming up sometime soon. But this is a wrong argument, because coin flips, at least in our model, are independent. The fact that these two happened to be tails doesn't change anything about our beliefs about what's going to be happening here.

So what's going to be happening to that person is they will be flipping independent coin flips. That person will also be flipping independent coin flips. And both of them wait until the first head occurs. They're facing an identical situation, starting from this time.

OK, now what's the probabilistic model of what this person is facing? The time until that person obtains heads for the first time is X . So this number of flips until they obtain heads for the first time is going to be X minus 2. So X is the total number until the first head. X minus 2 is the number of flips, starting from here.

Now what information do we have about that person? We have the information that their first two flips were tails. So we're given the information that X was bigger than 2. So the probabilistic model that describes this piece of the experiment is that it's going to take a random number of flips until the first head.

That number of flips, starting from here until the next head, is that number X minus 2. But we're given the information that this person has already wasted 2 coin flips. Now we argued that probabilistically, this person, this part of the experiment here is identical with that part of the experiment. So the PMF of this random variable, which is X minus 2, conditioned on this information, should be the same as that PMF that we have down there.

So the formal statement that I'm making is that this PMF here of X minus 2, given that X is bigger than 2, is the same as the PMF of X itself. What is this saying? Given that I tell you that you already did a few flips and they were failures, the remaining number of flips until the first head has the same geometric distribution as if you were starting from scratch. Whatever happened in the past, it happened, but has no bearing what's going to happen in the future. Remaining coin flips until a head has the same distribution, whether you're starting right now, or whether you had done some other stuff in the past.

So this is a property that we call the memorylessness property of the geometric distribution. Essentially, it says that whatever happens in the future is independent from whatever happened in the past. And that's true almost by definition, because we're assuming independent coin flips. Really, independence means that information about one part of the experiment has no bearing about what's going to happen in the other parts of the experiment.

The argument that I tried to give using the intuition of coin flips, you can make it formal by just manipulating PMFs formally. So this is the original PMF of X .

Suppose that you condition on the event that X is bigger than 3. This conditioning information, what it does is it tells you that this piece did not happen. You're conditioning just on this event.

When you condition on that event, what's left is the conditional PMF, which has the same shape as this one, except that it needs to be re-normalized up, so that the probabilities add up to one. So you take that picture, but you need to change the height of it, so that these terms add up to 1.

And this is the conditional PMF of X , given that X is bigger than 2. But we're talking here not about X . We're talking about the remaining number of heads. Remaining number of heads is X minus 2.

If we have the PMF of X , can we find the PMF of X minus 2? Well, if X is equal to 3, that corresponds to X minus 2 being equal to 1. So this probability here should be equal to that probability.

The probability that X is equal to 4 should be the same as the probability that X minus 2 is equal to 2. So basically, the PMF of X minus 2 is the same as the PMF of X , except that it gets shifted by these 2 units.

So this way, we have formally derived the conditional PMF of the remaining number of coin tosses, given that the first two flips were tails. And we see that it's exactly the same as the PMF that we started with. And so this is the formal proof of this statement here. So it's useful here to digest both these formal statements and understand it and understand the notation that is involved here, but also to really appreciate the intuitive argument what this is really saying.

OK, all right, so now we want to use this observation, this memorylessness, to eventually calculate the expected value for a geometric random variable. And the way we're going to do it is by using a divide and conquer tool, which is an analog of what we have already seen sometime before. Remember our story that there's a number of possible scenarios about the world? And there's a certain event, B , that can happen under any of these possible scenarios.

And we have the total probability theory. And that tells us that, to find the probability of this event, B , you consider the probabilities of B under each scenario. And you weigh those probabilities according to the probabilities of the different scenarios that we have. So that's a formula that we already know and have worked with.

What's the next step? Is it something deep? No, it's just translation in different notation.

This is the exactly same formula, but with PMFs. The event that capital X is equal to little x can happen in many different ways. It can happen under either scenario.

And within each scenario, you need to use the conditional probabilities of that event, given that this scenario has occurred. So this formula is identical to that one, except that we're using conditional PMFs, instead of conditional probabilities. But conditional PMFs, of course, are nothing but conditional probabilities anyway. So nothing new so far.

Then what I do is to take this formula here and multiply both sides by X and take the sum over all X 's. What do we get on this side? We get the expected value of X .

What do we get on that side? Probability of A_1 . And then here, sum over all X 's of X times P . That's, again, the same calculation we have when we deal with expectations, except that, since here, we're dealing with conditional probabilities, we're going to get the conditional expectation.

And this is the total expectation theorem. It's a very useful way for calculating expectations using a divide and conquer method.

We figure out the average value of X under each one of the possible scenarios. The overall average value of X is a weighted linear combination of the expected values of X in the different scenarios where the weights are chosen according to the different probabilities.

OK, and now we're going to apply this to the case of a geometric random variable. And we're going to divide and conquer by considering separately the two cases where the first toss was heads, and the other case where the first toss was tails. So the expected value of X is the probability that the first toss was heads, so that X is equal to 1, and the expected value if that happened.

What is the expected value of X , given that X is equal to 1? If X is known to be equal to 1, then X becomes just a number. And the expected value of a number is the number itself. So this first line here is the probability of heads in the first toss times the number 1.

So the probability that X is bigger than 1 is 1 minus P . And then we need to do something about this conditional expectation. What is it?

I can write it in, perhaps, a more suggested form, as expected the value of X minus 1, given that X minus 1 is bigger than 1. Ah. OK, X bigger than 1 is the same as X minus 1 being positive, this way. X minus 1 is positive plus 1.

What did I do here? I added and subtracted 1. Now what is this? This is the expected value of the remaining coin flips, until I obtain heads, given that the first one was tails.

It's the same story that we were going through down there. Given that the first coin flip was tails doesn't tell me anything about the future, about the remaining coin flips. So this expectation should be the same as the expectation faced by a person who was starting just now. So this should be equal to the expected value of X itself. And then we have the plus 1 that's come from there, OK?

Remaining coin flips until a head, given that I had a tail yesterday, is the same as expected number of flips until heads for a person just starting now and wasn't doing anything yesterday. So the fact that they I had a coin flip yesterday doesn't change my beliefs about how long it's going to take me until the first head. So once we believe that relation, than we plug this here. And this red term becomes expected value of X plus 1.

So now we didn't exactly get the answer we wanted, but we got an equation that involves the expected value of X . And it's the only unknown in that equation. Expected value of X equals to P plus (1 minus P) times this expression. You solve this equation for expected value of X , and you get the value of $1/P$.

The final answer does make intuitive sense. If P is small, heads are difficult to obtain. So you expect that it's going to take you a long time until you see heads for the first time. So it is definitely a reasonable answer.

Now the trick that we used here, the divide and conquer trick, is a really nice one. It gives us a very good shortcut in this problem. But you must definitely spend some time making sure you understand why this expression here is the same as that expression there.

Essentially, what it's saying is that, if I tell you that X is bigger than 1, that the first coin flip was tails, all I'm telling you is that that person has wasted a coin flip, and they are starting all over again. So they've wasted 1 coin flip. And they're starting all over again. If I tell you that the first flip was tails, that's the only information that I'm basically giving you, a wasted flip, and then starts all over again.

All right, so in the few remaining minutes now, we're going to quickly introduce a few new concepts that we will be playing with in the next ten days or so. And you will get plenty of opportunities to manipulate them. So here's the idea.

A typical experiment may have several random variables associated with that experiment. So a typical student has height and weight. If I give you the PMF of height, that tells me something about distribution of heights in the class. I give you the PMF of weight, it tells me something about the different weights in this class.

But if I want to ask a question, is there an association between height and weight, then I need to know a little more how height and weight relate to each other. And the PMF of height individuality and PMF of weight just by itself do not tell me anything about those relations. To be able to say something about those relations, I need to know something about joint probabilities, how likely is it that certain X 's go together with certain Y 's. So these probabilities, essentially, capture associations between these two random variables. And it's the information I would need to have to do any kind of statistical study that tries to relate the two random variables with each other.

These are ordinary probabilities. This is an event. It's the event that this thing happens and that thing happens.

This is just the notation that we will be using. It's called the joint PMF. It's the joint Probability Mass Function of the two random variables X and Y looked at together, jointly. And it gives me the probability that any particular numerical outcome pair does happen.

So in the finite case, you can represent joint PMFs, for example, by a table. This particular table here would give you information such as, let's see, the joint PMF evaluated at 2, 3. This is the probability that X is equal to 3 and, simultaneously, Y is equal to 3. So it would be that number here. It's 4/20.

OK, what is a basic property of PMFs? First, these are probabilities, so all of the entries have to be non-negative. If you adopt the probabilities over all possible numerical pairs that you could get, of course, the total probability must be equal to 1. So that's another thing that we want.

Now suppose somebody gives me this model, but I don't care about Y's. All I care is the distribution of the X's. So I'm going to find the probability that X takes on a particular value.

Can I find it from the table? Of course, I can. If you ask me what's the probability that X is equal to 3, what I'm going to do is to add up those three probabilities together.

And those probabilities, taken all together, give me the probability that X is equal to 3. These are all the possible ways that the event X equals to 3 can happen. So we add these, and we get the 6/20.

What I just did, can we translate it to a formula? What did I do? I fixed the particular value of X. And I added up the values of the joint PMF over all the possible values of Y.

So that's how you do it. You take the joint. You take one slice of the joint, keeping X fixed, and adding up over the different values of Y.

The moral of this example is that, if you know the joint PMFs, then you can find the individual PMFs of every individual random variable. And we have a name for these. We call them the marginal PMFs.

We have the joint that talks about both together, and the marginal that talks about them one at the time. And finally, since we love conditional probabilities, we will certainly want to define an object called the conditional PMF.

So this quantity here is a familiar one. It's just a conditional probability. It's the probability that X takes on a particular value, given that Y takes a certain value.

For our example, let's take little y to be equal to 2, which means that we're conditioning to live inside this universe. This red universe here is the y equal to 2 universe. And these are the conditional probabilities of the different X's inside that universe.

OK, once more, just an exercise in notation. This is the chapter two version of the notation of what we were denoting this way in chapter one. The way to read this is that it's a conditional PMF having to do with two random variables, the PMF of X conditioned on information about Y. We are fixing a particular value of capital Y, that's the value on which we are conditioning. And we're looking at the probabilities of the different X's.

So it's really a function of two arguments, little x and little y. But the best way to think about it is to fix little y and think of it as a function of X. So I'm fixing little y here, let's say, to y equal to 2. So I'm considering only this.

And now, this quantity becomes a function of little x. For the different little x's, we're going to have different conditional probabilities. What are those conditional probabilities?

OK, conditional probabilities are proportional to original probabilities. So it's going to be those numbers, but scaled up. And they need to be scaled so that they add up to 1.

So we have 1, 3 and 1. That's a total of 5. So the conditional PMF would have the shape zero, $1/5$, $3/5$, and $1/5$. This is the conditional PMF, given a particular value of Y. It has the same shape as those numbers, where by shape, I mean try to visualize a bar graph.

The bar graph associated with those numbers has exactly the same shape as the bar graph associated with those numbers. The only thing that has changed is the scaling. Big moral, let me say in different words, the conditional PMF, given a particular value of Y, is just a slice of the joint PMF where you maintain the same shape, but you rescale the numbers so that they add up to 1.

Now mathematically, of course, what all of this is doing is it's taking the original joint PDF and it rescales it by a certain factor. This does not involve X, so the shape, is a function of X, has not changed. We're keeping the same shape as a function of X, but we divide by a certain number. And that's the number that we need, so that the conditional probabilities add up to 1.

Now where does this formula come from? Well, this is just the definition of conditional probabilities. Probability of something conditioned on something else is the probability of both things happening, the intersection of the two divided by the probability of the conditioning event.

And last remark is that, as I just said, conditional probabilities are nothing different than ordinary probabilities. So a conditional PMF must sum to 1, no matter what you are conditioning on. All right, so this was sort of quick introduction into our new notation. But you get a lot of practice in the next days to come.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013 Transcript – Tutorial: The Coupon Collector Problem

In this exercise, we'll be looking at a problem, also known as the coupons collector's problem. We have a set of K coupons, or grades in our case. And each time slot we're revealed with one random grade. And we'd like to know how long it would take for us to collect all K grades. In our case, K is equal to 6.

Now the key to solving the problem is essentially twofolds. First, we'll have to find a way to intelligently define sequence random variables that captured, essentially, stopping time of this process. And then we'll employ the idea of linearity of expectations in breaking down this value in simpler terms. So let's get started.

We'll define Y_i as the number of papers till we see the i -th new grade. What does that mean? Well, let's take a look at an example. Suppose, here we have a timeline from no paper yet, first paper, second paper, third paper, so on, and so forth. Now, if we got grade A on the first slot, grade A minus on second slot, A again on the third slot, let's say there's a fourth's slot, we got B.

According to this process, we see that Y_1 is always 1, because whatever we got on the first slot will be a new grade. Now, Y_2 is 2, because the second paper is, again, a new grade. On the third paper we got a grade, which is the same as the first grade. So that would not count as any Y_i . And the third time we saw new grade would now be paper four.

According to this notation, we're interested in knowing what is the expected value of E of Y_6 , which is the time it takes to receive all six grades. So so far this notation isn't really helping us in solving the problem, but kind of just staying a different way. It turns out, it's much easier to look at the following variable derived from the Y_i s.

We'll define X_i as the difference between Y_i plus 1 minus Y_i . And in [? words, ?] it says, X_i is a number of papers you need until you see the i plus 1-th new grade, after you have received i new grade so far. So in this case, X_1 will be if we call 0, Y_0 , will be the difference between Y_1 and Y_0 , which is always 1-- that's X_1 .

And the difference between these two will be X_2 . And the difference between Y_3 and Y_2 -- Sorry. It should be $Y_2 - Y_1$, 1, 2, and so on. OK?

Through this notation we see that Y_6 now can be written as the summation of i equal to 0, 2, 5, X_i . So all I did was to break down i_6 into a sequence of summation of the differences, like $Y_6 - Y_5$, $Y_5 - Y_4$, and so on. It turns out, this expression will be very useful. OK.

So now that we have the two variables Y and X , let's see if it will be easier to look at the distribution of X in studying this process. Let's say, we have seen a new grade so far-- one. How many trials would it take for us to see the second new grade?

It turns out it's not that hard. In this case, we know there is a total of six grades, and we have seen one of them. So that leaves us five more grades that we'll potentially see. And therefore, on any random trial after that, there is a probability of 5 over 6 that we'll see a new grade. And hence, we know that X_1 has a distribution geometric with a success probability, or a parameter, 5/6.

Now, more generally, if we extend this idea further, we see that X_i will have a geometric distribution of parameter 6 minus i over 6. And this is due to the fact that so far we have already seen i new grades. And that will be the success probability of seeing a further new grade.

So from this expression, we know that the expected value of X_i will simply be the inverse of the parameter of the geometric distribution, which is 6 over 6 minus i or 6 times 1 over 6 minus i . And now we're ready to compute a final answer.

So from this expression we know expected value of Y_6 is equal to the expected value of sum of i equal to 0 to 5 X_i . And by the linearity of expectation, we can pull out the sum and write it as 2, 5 expected value of X_i .

Now, since we know that expected value of X_i is the following expression. We see that this term is equal to 6 times expected value of i equals 0, 5, 1 over 6 minus i . Or written in the other way this is equal to 6 times i equal to 0, 2, 5. In fact, 1, 2, 5, 1 over i .

And all I did here was to, essentially, change the variable, so that these two summations contained exactly the same terms. And this will give us the answer, which is 14.7. Now, more generally, we can see that there's nothing special about number 6 here. We could have substituted 6 with a number, let's say, K .

And then we'll get E of Y_K , let's say, there's more than six labels. And this will give us K times summation i equal to 1, so K minus 1, 1 over i . Interestingly, it turns out this quantity has an [? asymptotic ?] expression that, essentially, is roughly equal to K times the natural logarithm of K . And this is known as the scaling [? la ?] for the coupon collector's problem that says, essentially, takes about K times [? la ?] K many trials until we collect all K coupons. And that'll be the end of the problem. See you next time.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

LECTURE 6

- **Readings:** Sections 2.4-2.6

Lecture outline

- Review: PMF, expectation, variance
- Conditional PMF
- Geometric PMF
- Total expectation theorem
- Joint PMF of two random variables

Review

- Random variable X : function from sample space to the real numbers
- PMF (for discrete random variables):
 $p_X(x) = P(X = x)$
- Expectation:

$$E[X] = \sum_x x p_X(x)$$

$$E[g(X)] = \sum_x g(x) p_X(x)$$

$$E[\alpha X + \beta] = \alpha E[X] + \beta$$

- $E[X - E[X]] =$

$$\begin{aligned} \text{var}(X) &= E[(X - E[X])^2] \\ &= \sum_x (x - E[X])^2 p_X(x) \\ &= E[X^2] - (E[X])^2 \end{aligned}$$

Standard deviation: $\sigma_X = \sqrt{\text{var}(X)}$

Random speed

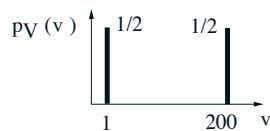
- Traverse a 200 mile distance at constant but random speed V



- $d = 200$, $T = t(V) = 200/V$
- $E[V] =$
- $\text{var}(V) =$
- $\sigma_V =$

Average speed vs. average time

- Traverse a 200 mile distance at constant but random speed V

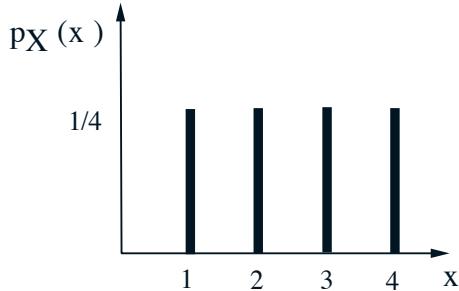


- time in hours = $T = t(V) =$
- $E[T] = E[t(V)] = \sum_v t(v)p_V(v) =$
- $E[TV] = 200 \neq E[T] \cdot E[V]$
- $E[200/V] = E[T] \neq 200/E[V]$.

Conditional PMF and expectation

- $p_{X|A}(x) = P(X = x | A)$

- $E[X | A] = \sum_x x p_{X|A}(x)$



- Let $A = \{X \geq 2\}$

$$p_{X|A}(x) =$$

$$E[X | A] =$$

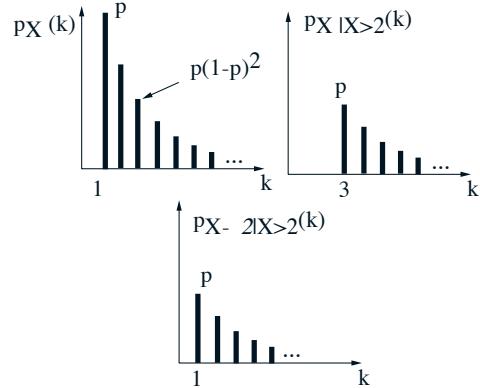
Geometric PMF

- X : number of independent coin tosses until first head

$$p_X(k) = (1-p)^{k-1}p, \quad k = 1, 2, \dots$$

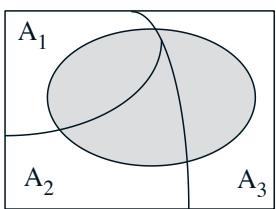
$$E[X] = \sum_{k=1}^{\infty} kp_X(k) = \sum_{k=1}^{\infty} k(1-p)^{k-1}p$$

- Memoryless property: Given that $X > 2$, the r.v. $X - 2$ has same geometric PMF



Total Expectation theorem

- Partition of sample space into disjoint events A_1, A_2, \dots, A_n



$$P(B) = P(A_1)P(B | A_1) + \dots + P(A_n)P(B | A_n)$$

$$p_X(x) = P(A_1)p_{X|A_1}(x) + \dots + P(A_n)p_{X|A_n}(x)$$

$$E[X] = P(A_1)E[X | A_1] + \dots + P(A_n)E[X | A_n]$$

- Geometric example:

$A_1 : \{X = 1\}, \quad A_2 : \{X > 1\}$

$$E[X] = P(X = 1)E[X | X = 1] + P(X > 1)E[X | X > 1]$$

- Solve to get $E[X] = 1/p$

Joint PMFs

- $p_{X,Y}(x,y) = P(X = x \text{ and } Y = y)$

		y				
		1	2	3	4	
		1	1/20	2/20	2/20	
		2		1/20	3/20	1/20
		3	2/20	4/20	1/20	2/20
		4				

- $\sum_x \sum_y p_{X,Y}(x,y) =$

- $p_X(x) = \sum_y p_{X,Y}(x,y)$

- $p_{X|Y}(x | y) = P(X = x | Y = y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}$

- $\sum_x p_{X|Y}(x | y) =$

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

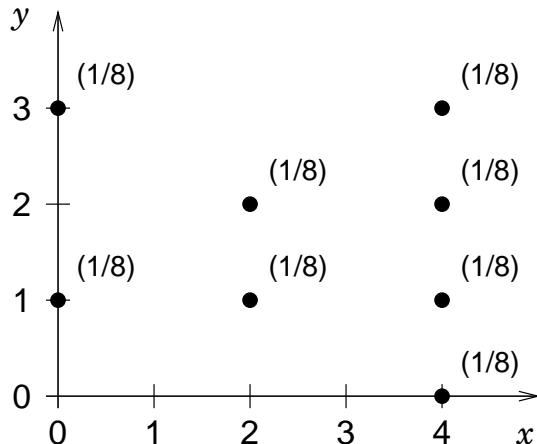
For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Recitation 6
September 28, 2010

1. Consider an experiment in which a fair four-sided die (with faces labeled 0, 1, 2, 3) is thrown once to determine how many times a fair coin is to be flipped. In the sample space of this experiment, random variables N and K are defined by

- N = the result of the die roll
 - K = the total number of heads resulting from the coin flips
- Determine and sketch $p_N(n)$
 - Determine and tabulate $p_{N,K}(n, k)$
 - Determine and sketch $p_{K|N}(k | 2)$
 - Determine and sketch $p_{N|K}(n | 2)$

2. Consider an outcome space comprising eight equally likely event points, as shown below:



- Which value(s) of x maximize(s) $\mathbf{E}[Y | X = x]$?
 - Which value(s) of y maximize(s) $\text{var}(X | Y = y)$?
 - Let $R = \min(X, Y)$. Prepare a neat, fully labeled sketch of $p_R(r)$,
 - Let A denote the event $X^2 \geq Y$. Determine numerical values for the quantities $\mathbf{E}[XY]$ and $\mathbf{E}[XY | A]$.
3. **Example 2.17. Variance of the geometric distribution.** You write a software program over and over, and each time there is probability p that it works correctly, independent of previous attempts. What is the variance of X , the number of tries until the program works correctly?

MIT OpenCourseWare
<http://ocw.mit.edu>

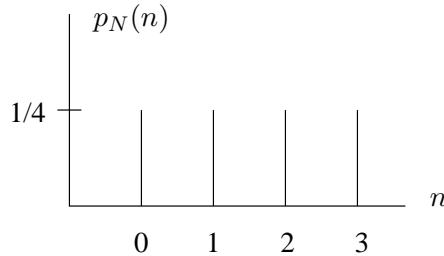
6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Recitation 6 Solutions
September 28, 2010

1. (a) The first part can be completed without reference to anything other than the die roll:



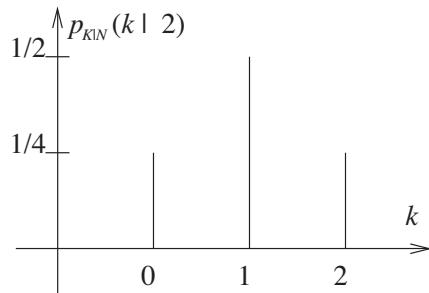
- (b) When $N = 0$, the coin is not flipped at all, so $K = 0$. When $N = n$ for $n \in \{1, 2, 3\}$, the coin is flipped n times, resulting in K with a distribution that is conditionally binomial. The binomial probabilities are all multiplied by $1/4$ because $p_N(n) = 1/4$ for $n \in \{0, 1, 2, 3\}$. The joint PMF $p_{N,K}(n, k)$ thus takes the following values and is zero otherwise:

	$k = 0$	$k = 1$	$k = 2$	$k = 3$
$n = 0$	1/4	0	0	0
$n = 1$	1/8	1/8	0	0
$n = 2$	1/16	1/8	1/16	0
$n = 3$	1/32	3/32	3/32	1/32

- (c) Conditional on $N = 2$, K is a binomial random variable. So we immediately see that

$$p_{K|N}(k|2) = \begin{cases} 1/4, & \text{if } k = 0, \\ 1/2, & \text{if } k = 1, \\ 1/4, & \text{if } k = 2, \\ 0, & \text{otherwise.} \end{cases}$$

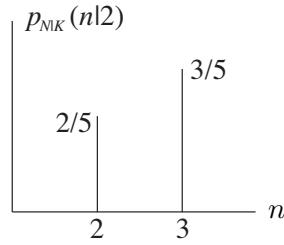
This is a normalized row of the table in the previous part.



- (d) To get $K = 2$ heads, there must have been at least 3 coin tosses, so only $N = 3$ and $N = 4$ have positive conditional probability given $K = 2$.

$$p_{N|K}(2|2) = \frac{\mathbf{P}(\{N = 2\} \cap \{K = 2\})}{\mathbf{P}(\{K = 2\})} = \frac{1/16}{1/16 + 1/32 + 1/32 + 1/32} = 2/5.$$

Similarly, $p_{N|K}(3|2) = 3/5$.



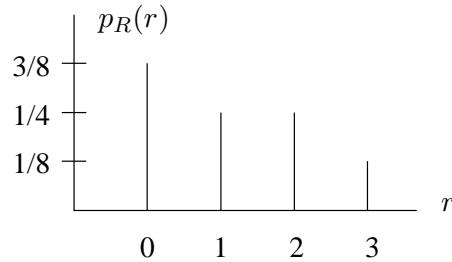
2. (a) $x = 0$ maximizes $\mathbf{E}[Y \mid X = x]$ since

$$\mathbf{E}[Y \mid X = x] = \begin{cases} 2, & \text{if } x = 0, \\ 3/2, & \text{if } x = 2, \\ 3/2, & \text{if } x = 4, \\ \text{undefined,} & \text{otherwise.} \end{cases}$$

- (b) $y = 3$ maximizes $\text{var}(X \mid Y = y)$ since

$$\text{var}(X \mid Y = y) = \begin{cases} 0, & \text{if } y = 0, \\ 8/3, & \text{if } y = 1, \\ 1, & \text{if } y = 2, \\ 4, & \text{if } y = 3, \\ \text{undefined,} & \text{otherwise.} \end{cases}$$

(c)



- (d) By traversing the points top to bottom and left to right, we obtain

$$\mathbf{E}[XY] = \frac{1}{8} (0 \cdot 3 + 4 \cdot 3 + 2 \cdot 2 + 4 \cdot 2 + 0 \cdot 1 + 2 \cdot 1 + 4 \cdot 1 + 4 \cdot 0) = \frac{15}{4}.$$

Conditioning on A removes the point masses at $(0, 1)$ and $(0, 3)$. The conditional probability of each of the remaining point masses is thus $1/6$, and

$$\mathbf{E}[XY \mid A] = \frac{1}{6} (4 \cdot 3 + 2 \cdot 2 + 4 \cdot 2 + 2 \cdot 1 + 4 \cdot 1 + 4 \cdot 0) = 5.$$

3. See the textbook, Example 2.17, pages 105–106.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Tutorial 3
September 30/October 1, 2010

1. Let X and Y be independent random variables. Random variable X has mean μ_X and variance σ_X^2 , and random variable Y has mean μ_Y and variance σ_Y^2 . Let $Z = 2X - 3Y$. Find the mean and variance of Z in terms of the means and variances of X and Y .
2. Problem 2.40, page 133 in the text.
A particular professor is known for his arbitrary grading policies. Each paper receives a grade from the set $\{A, A-, B+, B, B-, C+\}$, with equal probability, independently of other papers. How many papers do you expect to hand in before you receive each possible grade at least once?
3. The joint PMF of the random variables X and Y is given by the following table:

$y = 3$	c	c	$2c$
$y = 2$	$2c$	0	$4c$
$y = 1$	$3c$	c	$6c$
	$x = 1$	$x = 2$	$x = 3$

- (a) Find the value of the constant c .
- (b) Find $p_Y(2)$.
- (c) Consider the random variable $Z = YX^2$. Find $\mathbf{E}[Z \mid Y = 2]$.
- (d) Conditioned on the event that $X \neq 2$, are X and Y independent? Give a one-line justification.
- (e) Find the conditional variance of Y given that $X = 2$.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Tutorial 3: Solutions

1. In general we have that $\mathbf{E}[aX + bY + c] = a\mathbf{E}[X] + b\mathbf{E}[Y] + c$. Therefore,

$$\mathbf{E}[Z] = 2 \cdot \mathbf{E}[X] - 3 \cdot \mathbf{E}[Y].$$

For the case of independent random variables, we have that if $Z = a \cdot X + b \cdot Y$, then

$$\text{var}(Z) = a^2 \cdot \text{var}(X) + b^2 \cdot \text{var}(Y).$$

Therefore, $\text{var}(Z) = 4 \cdot \text{var}(X) + 9 \cdot \text{var}(Y)$.

2. See online solutions.
3. (a) We can find c knowing that the probability of the entire sample space must equal 1.

$$\begin{aligned} 1 &= \sum_{x=1}^3 \sum_{y=1}^3 p_{X,Y}(x,y) \\ &= c + c + 2c + 2c + 4c + 3c + c + 6c \\ &= 20c \end{aligned}$$

Therefore, $c = \frac{1}{20}$.

$$(b) p_Y(2) = \sum_{x=1}^3 p_{X,Y}(x,2) = 2c + 0 + 4c = 6c = \frac{3}{10}.$$

$$(c) Z = YX^2$$

$$\begin{aligned} \mathbf{E}[Z \mid Y = 2] &= \mathbf{E}[YX^2 \mid Y = 2] \\ &= \mathbf{E}[2X^2 \mid Y = 2] \\ &= 2\mathbf{E}[X^2 \mid Y = 2] \end{aligned}$$

$$p_{X|Y}(x \mid 2) = \frac{p_{X,Y}(x,2)}{p_Y(2)}.$$

Therefore,

$$p_{X|Y}(x \mid 2) = \begin{cases} \frac{1/10}{3/10} = \frac{1}{3} & \text{if } x = 1 \\ \frac{1/5}{3/10} = \frac{2}{3} & \text{if } x = 3 \\ 0 & \text{otherwise} \end{cases}$$

$$\begin{aligned} \mathbf{E}[Z \mid Y = 2] &= 2 \sum_{x=1}^3 x^2 p_{X|Y}(x \mid 2) \\ &= 2 \left((1^2) \cdot \frac{1}{3} + (3^2) \cdot \frac{2}{3} \right) \\ &= \frac{38}{3} \end{aligned}$$

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
 (Fall 2010)

(d) Yes. Given $X \neq 2$, the distribution of X is the same given $Y = y$.

$$\mathbf{P}(X = x | Y = y, X \neq 2) = \mathbf{P}(X = x | X \neq 2).$$

For example,

$$\mathbf{P}(X = 1 | Y = 1, X \neq 2) = \mathbf{P}(X = 1 | Y = 3, X \neq 2) = \mathbf{P}(X = 1 | X \neq 2) = \frac{1}{3}$$

(e) $p_{Y|X}(y | 2) = \frac{p_{X,Y}(2,y)}{p_X(2)}$.

$$p_X(2) = \sum_{y=1}^3 p_{X,Y}(2,y) = c + 0 + c = 2c = \frac{1}{10}.$$

Therefore,

$$p_{Y|X}(y | 2) = \begin{cases} \frac{1/20}{1/10} = \frac{1}{2} & \text{if } y = 1 \\ \frac{1/20}{1/10} = \frac{1}{2} & \text{if } y = 3 \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbf{E}[Y^2 | X = 2] = \sum_{y=1}^3 y^2 p_{Y|X}(y | 2) = (1^2) \cdot \frac{1}{2} + (3^2) \cdot \frac{1}{2} = 5.$$

$$\mathbf{E}[Y | X = 2] = \sum_{y=1}^3 y p_{Y|X}(y | 2) = 1 \cdot \frac{1}{2} + 3 \cdot \frac{1}{2} = 2.$$

$$\text{var}(Y | X = 2) = \mathbf{E}[Y^2 | X = 2] - \mathbf{E}[Y | X = 2]^2 = 5 - 2^2 = 1.$$

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013 Transcript – Recitation: Joint Probability Mass Function (PMF) Drill 2

Hey, guys. Welcome back. Today, we're going to do another fun problem, which is a drill problem on joint PMFs. And the goal is that you will feel more comfortable by the end of this problem, manipulating joint PMFs. And we'll also review some ideas about independents in the process.

So just to go over what I've drawn here, we are given an xy plane. And we're told what the PMF is. And it's plotted for you here.

What these stars indicate is simply that there is a value there. But we don't know what it is. It could be anything between 0 and 1. And so we're given this list of questions. And we're just going to work through them linearly together.

So we start off pretty simply. We want to compute, in part a, the probability that x takes on a value of 1. So for those of you who like formulas, I'm going to use the formula, which is usually referred to as marginalization.

So the marginal over x is given by summing over the joint. So here we are interested in the probability that x is 1. So I'm just going to freeze the value of 1 here. And we sum over y .

And in particular, 1, 2, and 3. So carrying this out, this is the P_{xy} of 1, 1, plus P_{xy} of 1, 2, plus P_{xy} 1, 3. And this, of course, reading from the graph, is $1/12$ plus $2/12$ plus $1/12$, which is equal to $4/12$, or $1/3$.

So now you guys know the formula. Hopefully you'll remember the term marginalization. But I want to point out that intuitively you can come up with the answer much faster.

So the probability that x is equal to 1 is the probability that this dot happens or this dot happens or this dot happens. Now, these dots, or outcomes, they're disjoint. So you can just sum the probability to get the probability of one of these things happening.

So it's the same computation. And you'll probably get there a little bit faster. So we're done with a already, which is great.

So for part b, conditioning on x is equal to 1, we want to sketch the PMF of y . So if x is equal to 1 we are suddenly living in this universe. y can take values of 1, 2, or 3 with these relative frequencies.

So let's draw this here. So this is y . I said, already, y can take on a value of 1. y can take on a value of 2. Or it can take on a value of 3. And we're plotting here, $P(y \text{ given } x, y)$, conditioned on x is equal to 1.

OK, so what I mean by preserving the relative frequencies is that in unconditional world this dot is twice as likely to happen as either this dot or this dot. And that relative likelihood remains the same after conditioning. And the reason why we have to change these values is because they have to sum to 1. So in other words, we have to scale them up.

So you can use a formula. But again, I'm here to show you faster ways of thinking about it. So my little algorithm for figuring out conditional PMFs is to take the numerators-- so 1, 2, and 1-- and sum them. So here that gives us 4.

And then to preserve the relative frequency, you actually keep the same numerators but divide it by the sum, which you just computed. So I'm going fast. I'll review in a second. But this is what you will end up getting.

So to recap, I did 1 plus 2 plus 1, which is 4, to get these denominators. And so I skipped a step here. This is really $2/4$, which is $1/2$, obviously.

So you add these guys to get 4. And then you keep the numerators and just divide them by 4. So $1/4$, $2/4$, which is $1/2$ and $1/4$. And that's what we mean by preserving the relative frequency. Except so this thing now sums to 1, which is what we want.

OK, so we're done with part b. Part c actually follows almost immediately from part b. In part c we're interested in computing the conditional expectation of y given that x is equal to 1.

So we've already done most of the legwork because we have the conditional PMF that we need. And so expectation, you guys have calculated a bunch of these by now. So I'm just going to appeal to your intuition and to symmetry.

Expectation acts like center of mass. This is a symmetrical distribution of mass. And so the center is right here at 2. So this is simply 2. And if that went too fast, just convince yourselves. Use the normal formula for expectations. And your answer will agree with ours.

OK, so d is a really cool question. Because you can do a lot of math. Or you can think and ask yourself, at the most fundamental level, what is independent? And if you think that way you'll come to the answer very easily.

So essentially, I rephrased this to truncate it from the problem statement that you guys are reading. But the idea is that these stars are unknown probability masses. And this question is asking can you figure out a way of assigning numbers between 0 and 1 to these values such that you end up with a valid probability mass function, so everything sums to 1 and such that x and y are independent?

So it seems hard a priori. But let's think about it a bit. And in the meantime I'm going to erase this so I have more room. What does it mean for x and y to be independent?

Well, it means that they don't, essentially, have information about each other. So if I tell you something about x and if x and y are independent, your belief about y shouldn't change. In other words, if you're a rational person, x shouldn't change your belief about y .

So let's look more closely at this diagram. Now, the number 0 should be popping out to you. Because this essentially means that the 0.31 can't happen. Or it happens with 0 probability.

So let's say fix x equal to 3. If you condition on x is equal to 3, as I just said, this outcome can't happen. So y could only take on values of 2 or 3.

However, if you condition on x is equal to 1, y could take on a value of 1 with probability 1/4 as we computed in the other problem. It could take on a value of 2 with probability of 1/2. Or it could take on a value of 3 with probability 1/4.

So these are actually very different cases, right? Because if you observe x is equal to 3 y can only be 2 or 3. But if you observe x is equal to 1, y can be 1, 2, or 3.

So actually, x , no matter what values these stars have on, x always tells you something about y . Therefore, the answer to this, part d, is no. So let's put a no with an exclamation point. So I like that problem a lot. And hopefully it clarifies independents for you guys.

So parts e and f, we're going to be thinking about independents again. To go over what the problem statement gives you, we defined this event, b , which is the event that x is less than or equal to 2 and y is less than or equal to 2.

So let's get some colors. So do bright pink. So that means we're essentially living in this world. There's only those four dots. And we're also told a very important piece of information that conditions on B . x and y are conditionally independent.

OK, so part e, now that we have this. And by the way, these two assumptions apply to both parts e and part f. So in part e, we want to find out P_{xy} of 2, 2. Or in English, what is the probability that x takes on a value of 2 and y takes on a value of 2?

So determine the value of this star. And the whole trick here is that the possible values that this star could take on are constrained by the fact that we need to make sure that x and y are conditionally independent given B .

So my claim is that if two random variables are independent and you condition on one of them, say we condition on x . If you condition on different values of x , the relative frequencies of y should be the same. So here, the relative frequency, condition on x is equal to 1. The relative frequencies of y are 2 to 1.

This outcome is twice as likely to happen as this one. If we condition on 2 this outcome needs to be twice as likely to happen as this outcome. If they weren't, x would tell you information about y. Because you would know that the distribution over 2 and 1 would be different. OK?

So because the relative frequencies have to be the same and $2/12$ is 2 times $1/12$ this guy must also be 2 times $2/12$. So that gives us our answer for part e. Let me write up here. Part e, we need $P_{xy}(2, 2)$ to be equal to $4/12$.

And again, the way we got this is simply we need x and y to be conditionally independent given B. And if this were anything other than 4 the relative frequency of y is equal to 2 to 1 would be different from over here. So here condition on x is equal to 1. The outcome, y is equal to 2 is twice as likely as x is equal to 1.

Here, if we put a value of $4/12$ and you condition on x is equal to 2, the outcome y is equal to 2 is still twice as likely as the outcome y is equal to 1. And if you put any other number there the relative frequencies would be different. So x would be telling you something about y.

So there would not be independent condition on B. OK, that was a mouthful. But hopefully you guys have it now.

And lastly, we have part f, which follows pretty directly from part e. So we were still in the unconditional universe. In part e, we were figuring out what's the value of star in the whole unconditional universe?

Now, in part f, we want the value of star in the conditional universe where B occurred. So let's come over here and plot a new graph so we don't confuse ourselves. So we have xy.

x can be 1 or 2. Y could be 1 or 2. So we have a plot that looks something like this.

And so again, same argument as before. Let me just fill this in. From part e, we have that this is $4/12$. And we're going to use my algorithm again.

So in the conditional world, the relative frequencies of these four dots should be the same. But you need to scale them up so that if you sum over all of them the probability sums to 1. So you have a valid PMF.

So my algorithm from before was to add up all the numerators. So 1 plus 2 plus 4 plus 2 gives you 9. And then to preserve the relative frequency you keep the same numerator.

So here we had a numerator of 1. That becomes $1/9$. Here we had a numerator of 2.

This becomes $2/9$. Here we had a numerator of 4. That becomes $4/9$. Here we had a numerator of 2, so $2/9$.

And indeed, the relative frequencies are preserved. And they all sum to 1. So our answer for part f-- let's box it here-- is that $P_{xy|B} = \frac{2}{9}$, 2 conditioned on B is equal to 4/9, is just that guy. So we're done.

Hopefully that wasn't too painful. And this is a good drill problem, because we got more comfortable working with PMFs, joint PMFs. We went over marginalization. We went over conditioning.

We went over into independents. And I also gave you this quick algorithm for figuring out what conditional PMFs are if you don't want to use the formulas. Namely, you sum all of the numerators to get a new denominator and then divide all the old numerators by the new denominator you computed. So I hope that was helpful. I'll see you next time.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013

Transcript – Lecture 7

The following content is provided under a Creative Commons license. Your support will help MIT OpenCourseWare continue to offer high-quality educational resources for free. To make a donation or view additional materials from hundreds of MIT courses, visit MIT OpenCourseWare at ocw.mit.edu.

PROFESSOR: OK, good morning. So today, we're going to have a fairly packed lecture. We are going to conclude with chapter two, discrete random variables. And we will be talking mostly about multiple random variables. And this is also the last lecture as far as quiz one is concerned. So it's going to cover the material until today, and of course the next recitation and tutorial as well.

OK, so we're going to review quickly what we introduced at the end of last lecture, where we talked about the joint PMF of two random variables. We're going to talk about the case of more than two random variables as well. We're going to talk about the familiar concepts of conditioning and independence, but applied to random variables instead of events. We're going to look at the expectations once more, talk about a few properties that they have, and then solve a couple of problems and calculate a few things in somewhat clever ways.

So the first point I want to make is that, to a large extent, whatever is happening in our chapter on discrete random variables is just an exercise in notation. There is stuff and concepts that you are already familiar with-- probabilities, probabilities of two things happening, conditional probabilities. And all that we're doing, to some extent, is rewriting those familiar concepts in new notation. So for example, this is the joint PMF of two random variable. It gives us, for any pair or possible values of those random variables, the probability that that pair occurs simultaneously. So it's the probability that simultaneously x takes that value, and y takes that other value.

And similarly, we have the notion of the conditional PMF, which is just a list of the -- condition of -- the various conditional probabilities of interest, conditional probability that one random variable takes this value given that the other random variable takes that value. Now, a remark about conditional probabilities. Conditional probabilities generally are like ordinary probabilities. You condition on something particular. So here we condition on a particular y . So think of little y as a fixed quantity. And then look at this as a function of x . So given that y , which we condition on, given our new universe, we're considering the various possibilities for x and the probabilities that they have.

Now, the probabilities over all x 's, of course, needs to add to 1. So we should have a relation of this kind. So they're just like ordinary probabilities over the different x 's in a universe where we are told the value of the random variable y . Now, how are these related? So we call these the marginal, these the joint, these the conditional. And there are some relations between these. For example, to find the marginal from the joint, it's pretty straightforward. The probability that x takes a particular value is the sum of the probabilities of all of the different ways that this particular value may occur.

What are the different ways? Well, it may occur together with a certain y , or together with some other y , or together with some other y . So you look at all the possible y 's that can go together with this x , and add the probabilities of all of those pairs for which we get this particular value of x . And then there's a relation between that connects these two probabilities with the conditional probability. And it's this relation. It's nothing new. It's just new notation for writing what we already know, that the probability of two things happening is the probability that the first thing happens, and then given that the first thing happens, the probability that the second one happened.

So how do we go from one to the other? Think of A as being the event that X takes the value, little x , and B being the event that Y takes the value, little y . So the joint probability is the probability that these two things happen simultaneously. It's the probability that X takes this value times the conditional probability that Y takes this value, given that X took that first value. So it's the familiar multiplication rule, but just transcribed in our new notation. So nothing new so far.

OK, why did we go through this exercise and this notation? It's because in the experiments where we're interested in the real world, typically there's going to be lots of uncertain quantities. There's going to be multiple random variables. And we want to be able to talk about them simultaneously. Okay. Why two and not more than two? How about three random variables? Well, if you understand what's going on in this slide, you should be able to kind of automatically generalize this to the case of multiple random variables.

So for example, if we have three random variables, X , Y , and Z , and you see an expression like this, it should be clear what it means. It's the probability that X takes this value and simultaneously Y takes that value and simultaneously Z takes that value. I guess that's an uppercase Z here, that's a lowercase z . And if I ask you to find the marginal of X , if I tell you the joint PMF of the three random variables and I ask you for this value, how would you find it? Well, you will try to generalize this relation here. The probability that x occurs is the sum of the probabilities of all events that make X to take that particular value.

So what are all the events? Well, this particular x can happen together with some y and some z . We don't care which y and z . Any y and z will do. So when we consider all possibilities, we need to add here over all possible values of y 's and z 's. So consider all triples, x, y, z . Fix x and consider all the possibilities for the remaining variables, y and z , add these up, and that gives you the marginal PMF of X . And then there's other things that you can do. This is the multiplication rule for two events.

We saw back in chapter one that there's a multiplication rule when you talk about more than two events. And you can write a chain of conditional probabilities. We can certainly do the same in our new notation. So let's look at this rule up here. Multiplication rule for three random variables, what does it say? The probability of three things happening simultaneously, X, Y, Z taking specific values, little x , little y , little z , that probability is the probability that the first thing happens, that X takes that value.

Given that X takes that value, we multiply it with the probability that Y takes also a certain value. And now, given that X and Y have taken those particular values, we multiply with a conditional probability that the third thing happens, given that the first two things happen. So this is just the multiplication rule for three events, which would be probability of A intersection B intersection C equals-- you know the rest of the formula. You just rewrite this formula in PMF notation. Probability of A intersection B intersection C is the probability of A , which corresponds to this term, times the probability of B given A , times the probability of C given A and B .

So what else is there that's left from chapter one that we can or should generalize to random variables? Well, there's the notion of independence. So let's define what independence means. Instead of talking about just two random variables, let's go directly to the case of multiple random variables. When we talked about events, things were a little complicated. We had a simple definition for independence of two events. Two events are independent if the probability of both is equal to the product of the probabilities. But for three events, it was kind of messy. We needed to write down lots of conditions.

For random variables, things in some sense are a little simpler. We only need to write down one formula and take this as the definition of independence. Three random variables are independent if and only if, by definition, their joint probability mass function factors out into individual probability mass functions. So the probability that all three things happen is the product of the individual probabilities that each one of these three things is happening. So independence means mathematically that you can just multiply probabilities to get to the probability of several things happening simultaneously.

So with three events, we have to write a huge number of equations, of equalities that have to hold. How can it be that with random variables we can only manage with one equality? Well, the catch is that this is not really just one equality. We require this to be true for every little x , y , and z . So in some sense, this is a bunch of conditions that are being put on the joint PMF, a bunch of conditions that we need to check. So this is the mathematical definition. What is the intuitive content of this definition? The intuitive content is the same as for events. Random variables are independent if knowing something about the realized values of some of these random variables does not change our beliefs about the likelihood of various values for the remaining random variables.

So independence would translate, for example, to a condition such as the conditional PMF of X , given y , should be equal to the marginal PMF of X . What is this saying? That you have some original beliefs about how likely it is for X to take this value. Now, someone comes and tells you that Y took on a certain value. This causes you, in principle, to revise your beliefs. And your new beliefs will be captured by the conditional PMF, or the conditional probabilities. Independence means that your revised beliefs actually will be the same as your original beliefs. Telling you information about the value of Y doesn't change what you expect for the random variable X .

Why didn't we use this definition for independence? Well, because this definition only makes sense when this conditional is well-defined. And this conditional is only well-defined if the events that Y takes on that particular value has positive probability. We cannot condition on

events that have zero probability, so conditional probabilities are only defined for y's that are likely to occur, that have a positive probability.

Now, similarly, with multiple random variables, if they're independent, you would have relations such as the conditional of X, given y and z, should be the same as the marginal of X. What is this saying? Again, that if I tell you the values, the realized values of random variables Y and Z, this is not going to change your beliefs about how likely x is to occur. Whatever you believed in the beginning, you're going to believe the same thing afterwards. So it's important to keep that intuition in mind, because sometimes this way you can tell whether random variables are independent without having to do calculations and to check this formula.

OK, so let's check our concepts with a simple example. Let's look at two random variables that are discrete, take values between one and four each. And this is a table that gives us the joint PMF. So it tells us the probability that X equals to 2 and Y equals to 1 happening simultaneously. It's an event that has probability 1/20. Are these two random variables independent? You can try to check a condition like this. But can we tell directly from the table?

If I tell you a value of Y, could that give you useful information about X? Certainly. If I tell you that Y is equal to 1, this tells you that X must be equal to 2. But if I tell you that Y was equal to 3, this tells you that, still, X could be anything. So telling you the value of Y kind of changes what you expect or what you consider possible for the values of the other random variable. So by just inspecting here, we can tell that the random variables are not independent.

Okay. What's the other concept we introduced in chapter one? We introduced the concept of conditional independence. And conditional independence is like ordinary independence but applied to a conditional universe where we're given some information. So suppose someone tells you that the outcome of the experiment is such that X is less than or equal to 2 and Y is larger than or equal to 3. So we are given the information that we now live inside this universe.

So what happens inside this universe? Inside this universe, our random variables are going to have a new joint PMF which is conditioned on the event that we were told that it has occurred. So let A correspond to this sort of event here. And now we're dealing with conditional probabilities. What are those conditional probabilities? We can put them in a table. So it's a two by two table, since we only have two possible values. What are they going to be?

Well, these probabilities show up in the ratios 1, 2, 2, and 4. Those ratios have to stay the same. The probabilities need to add up to one. So what should the denominators be since these numbers add up to nine? These are the conditional probabilities. So this is the conditional PMF in this example. Now, in this conditional universe, is x independent from y? If I tell you that y takes this value, so we live in this universe, what do you know about x? What you know about x is at this value is twice as likely as that value. If I condition on y taking this value, so we're living here, what do you know about x? What you know about x is that this value is twice as likely as that value.

So it's the same. Whether we live here or we live there, this x is twice as likely as that x. So the conditional PMF in this new universe, the conditional PMF of X given y, in the new universe is

the same as the marginal PMF of X, but of course in the new universe. So no matter what y is, the conditional PMF of X is the same. And that conditional PMF is 1/3 and 2/3. This is the conditional PMF of X in the new universe no matter what y occurs.

So Y does not give us any information about X, doesn't cause us to change our beliefs inside this little universe. And therefore the two random variables are independent. Now, the other way that you can verify that we have independence is to find the marginal PMFs of the two random variables. The marginal PMF of X, you find it by adding those two terms. You get 1/3. Adding those two terms, you get 2/3. Marginal PMF of Y, you find it, you add these two terms, and you get 1/3. And the marginal PMF of Y here is going to be 2/3.

And then you ask the question, is the joint the product of the marginals? And indeed it is. This times this gives you 1/9. This times this gives you 2/9. So the values in the table with the joint PMFs is the product of the marginal PMFs of X and Y in this universe, so the two random variables are independent inside this universe. So we say that they're conditionally independent. All right.

Now let's move to the new topic, to the new concept that we introduce in this chapter, which is the concept of expectations. So what are the things to know here? One is the general idea. The way to think about expectations is that it's something like the average value for random variable if you do an experiment over and over, and if you interpret probabilities as frequencies. So you get x's over and over with a certain frequency -- $P(x)$ -- a particular value, little x, gets realized. And each time that this happens, you get x dollars. How many dollars do you get on the average? Well, this formula gives you that particular average.

So first thing we do is to write down a definition for this sort of concept. But then the other things you need to know is how to calculate expectations using shortcuts sometimes, and what properties they have. The most important shortcut there is is that, if you want to calculate the expected value, the average value for a random variable, you do not need to find the PMF of that random variable. But you can work directly with the x's and the y's. So you do the experiment over and over. The outcome of the experiment is a pair (x,y) . And each time that a certain (x,y) happens, you get so many dollars.

So this fraction of the time, a certain (x,y) happens. And that fraction of the time, you get so many dollars, so this is the average number of dollars that you get. So what you end up, since it is the average, then that means that it corresponds to the expected value. Now, this is something that, of course, needs a little bit of mathematical proof. But this is just a different way of accounting. And it turns out we give you the right answer. And it's a very useful shortcut.

Now, when we're talking about functions of random variables, in general, we cannot speak just about averages. That is, the expected value of a function of a random variable is not the same as the function of the expected values. A function of averages is not the same as the average of a function. So in general, this is not true. But what it's important to know is to know the exceptions to this rule. And the important exceptions are mainly two. One is the case of linear functions of a random variable. We discussed this last time. So the expected value of temperature in Celsius is, you first find the expected value of temperature in Fahrenheit, and then you do the conversion to

Celsius. So whether you first average and then do the conversion to the new units or not, it shouldn't matter when you get the result.

The other property that turns out to be true when you talk about multiple random variables is that expectation still behaves linearly. So let X , Y , and Z be the score of a random student at each one of the three sections of the SAT. So the overall SAT score is X plus Y plus Z . This is the average score, the average total SAT score. Another way to calculate that average is to look at the first section of the SAT and see what was the average. Look at the second section, look at what was the average, and so the third, and add the averages. So you can do the averages for each section separately, add the averages, or you can find total scores for each student and average them.

So I guess you probably believe that this is correct if you talk just about averaging scores. Since expectations are just the variation of averages, it turns out that this is also true in general. And the derivation of this is very simple, based on the expected value rule. And you can look at it in the notes. So this is one exception, which is linearity. The second important exception is the case of independent random variables, that the product of two random variables has an expectation which is the product of the expectations. In general, this is not true. But for the case where we have independence, the expectation works out as follows. Using the expected value rule, this is how you calculate the expected value of a function of a random variable. So think of this as being your $g(X, Y)$ and this being your $g(\text{little } x, y)$.

So this is something that's generally true. Now, if we have independence, then the PMFs factor out, and then you can separate this sum by bringing together the x terms, bring them outside the y summation. And you find that this is the same as expected value of X times the expected value of Y . So independence is used in this step here. OK, now what if X and Y are independent, but instead of taking the expectation of X times Y , we take the expectation of the product of two functions of X and Y ? I claim that the expected value of the product is still going to be the product of the expected values.

How do we show that? We could show it by just redoing this derivation here. Instead of X and Y , we would have $g(X)$ and $h(Y)$, so the algebra goes through. But there's a better way to think about it which is more conceptual. And here's the idea. If X and Y are independent, what does it mean? X does not convey any information about Y . If X conveys no information about Y , does X convey information about $h(Y)$? No. If X tells me nothing about Y , nothing new, it shouldn't tell me anything about $h(Y)$.

Now, if X tells me nothing about h of $h(Y)$, could $g(X)$ tell me something about $h(Y)$? No. So the idea is that, if X is unrelated to Y , doesn't have any useful information, then $g(X)$ could not have any useful information for $h(Y)$. So if X and Y are independent, then $g(X)$ and $h(Y)$ are also independent. So this is something that one can try to prove mathematically, but it's more important to understand conceptually why this is so. It's in terms of conveying information.

So if X tells me nothing about Y , X cannot tell me anything about Y cubed, or X cannot tell me anything by Y squared, and so on. That's the idea. And once we are convinced that $g(X)$ and $h(Y)$ are independent, then we can apply our previous rule, that for independent random

variables, expectations multiply the right way. Apply the previous rule, but apply it now to these two independent random variables. And we get the conclusion that we wanted.

Now, besides expectations, we also introduced the concept of the variance. And if you remember the definition of the variance, let me write down the formula for the variance of aX . It's the expected value of the random variable that we're looking at minus the expected value of the random variable that we're looking at. So this is the difference of the random variable from its mean. And we take that difference and square it, so it's the squared distance from the mean, and then take expectations of the whole thing.

So when you look at that expression, you realize that a can be pulled out of those expressions. And because there is a squared, when you pull out the a , it's going to come out as an a -squared. So that gives us the rule for finding the variance of a scale or product of a random variable. The variance captures the idea of how wide, how spread out a certain distribution is. Bigger variance means it's more spread out.

Now, if you take a random variable and the constants to it, what does it do to its distribution? It just shifts it, but it doesn't change its width. So intuitively it means that the variance should not change. You can check that mathematically, but it should also make sense intuitively. So the variance, when you add the constant, does not change. Now, can you add variances in the way we added expectations? Does variance behave linearly? It turns out that not always. Here, we need a condition. It's only in special cases-- for example, when the two random variables are independent-- that you can add variances. The variance of the sum is the sum of the variances if X and Y are independent.

The derivation of this is, again, very short and simple. We'll skip it, but it's an important fact to remember. Now, to appreciate why this equality is not true always, we can think of some extreme examples. Suppose that X is the same as Y . What's going to be the variance of X plus Y ? Well, X plus Y , in this case, is the same as $2X$, so we're going to get 4 times the variance of X , which is different than the variance of X plus the variance of X .

So that expression would give us twice the variance of X . But actually now it's 4 times the variance of X . The other extreme would be if X is equal to $-Y$. Then the variance is the variance of the random variable, which is always equal to 0. Now, a random variable which is always equal to 0 has no uncertainty. It is always equal to its mean value, so the variance, in this case, turns out to be 0.

So in both of these cases, of course we have random variables that are extremely dependent. Why are they dependent? Because if I tell you something about Y , it tells you an awful lot about the value of X . There's a lot of information about X if I tell you Y , in this case or in that case. And finally, a short drill. If I tell you that the random variables are independent and you want to calculate the variance of a linear combination of this kind, then how do you argue? You argue that, since X and Y are independent, this means that X and $3Y$ are also independent. X has no information about Y , so X has no information about $-Y$. X has no information about $-Y$, so X should not have any information about $-3Y$. So X and $-3Y$ are independent.

So the variance of Z should be the variance of X plus the variance of $-3Y$, which is the variance of X plus 9 times the variance of Y . The important thing to note here is that no matter what happens, you end up getting a plus here, not a minus. So that's the sort of important thing to remember in this type of calculation. So this has been all concepts, reviews, new concepts and all that. It's the usual fire hose. Now let's use them to do something useful finally.

So let's revisit our old example, the binomial distribution, which counts the number of successes in independent trials of a coin. It's a biased coin that has a probability of heads, or probability of success, equal to p at each trial. Finally, we can go through the exercise of calculating the expected value of this random variable. And there's the way of calculating that expectation that would be the favorite of those people who enjoy algebra, which is to write down the definition of the expected value. We add over all possible values of the random variable, over all the possible k 's, and weigh them according to the probabilities that this particular k occurs. The probability that X takes on a particular value k is, of course, the binomial PMF, which is this familiar formula.

Clearly, that would be a messy and challenging calculation. Can we find a shortcut? There's a very clever trick. There's lots of problems in probability that you can approach really nicely by breaking up the random variable of interest into a sum of simpler and more manageable random variables. And if you can make it to be a sum of random variables that are just 0's or 1's, so much the better. Life is easier. Random variables that take values 0 or 1, we call them indicator variables. They indicate whether an event has occurred or not.

In this case, we look at each coin flip one at a time. For the i -th flip, if it resulted in heads or a success, we record it 1. If not, we record it 0. And then we look at the random variable. If we take the sum of the X_i 's, what is it going to be? We add one each time that we get a success, so the sum is going to be the total number of successes. So we break up the random variable of interest as a sum of really nice and simple random variables.

And now we can use the linearity of expectations. We're going to find the expectation of X by finding the expectation of the X_i 's and then adding the expectations. What's the expected value of X_i ? Well, X_i takes the value 1 with probability p , and takes the value 0 with probability $1-p$. So the expected value of X_i is just p . So the expected value of X is going to be just n times p . Because X is the sum of n terms, each one of which has expectation p , the expected value of the sum is the sum of the expected values. So I guess that's a pretty good shortcut for doing this horrendous calculation up there.

So in case you didn't realize it, that's what we just established without doing any algebra. Good. How about the variance of X , of X_i ? Two ways to calculate it. One is by using directly the formula for the variance, which would be -- let's see what it would be. With probability p , you get a 1. And in this case, you are so far from the mean. That's your squared distance from the mean. With probability $1-p$, you get a 0, which is so far away from the mean. And then you can simplify that formula and get an answer.

How about a slightly easier way of doing it. Instead of doing the algebra here, let me indicate the slightly easier way. We have a formula for the variance that tells us that we can find the variance

by proceeding this way. That's a formula that's generally true for variances. Why is this easier? What's the expected value of X_i squared? Backtrack. What is X_i squared, after all? It's the same thing as X_i . Since X_i takes value 0 and 1, X_i squared also takes the same values, 0 and 1. So the expected value of X_i squared is the same as the expected value of X_i , which is equal to p . And the expected value of X_i squared is p squared, so we get the final answer, p times $(1-p)$.

If you were to work through and do the cancellations in this messy expression here, after one line you would also get to the same formula. But this sort of illustrates that working with this formula for the variance, sometimes things work out a little faster. Finally, are we in business? Can we calculate the variance of the random variable X as well? Well, we have the rule that for independent random variables, the variance of the sum is the sum of the variances. So to find the variance of X , we just need to add the variances of the X_i 's. We have n X_i 's, and each one of them has variance p_n times $(1-p)$. And we are done.

So this way, we have calculated both the mean and the variance of the binomial random variable. It's interesting to look at this particular formula and see what it tells us. If you are to plot the variance of X as a function of p , it has this shape. And the maximum is here at $1/2$. p times $(1-p)$ is 0 when p is equal to 0. And when p equals to 1, it's a quadratic, so it must have this particular shape. So what does it tell us? If you think about variance as a measure of uncertainty, it tells you that coin flips are most uncertain when your coin is fair. When p is equal to $1/2$, that's when you have the most randomness.

And this is kind of intuitive. if on the other hand I tell you that the coin is extremely biased, p very close to 1, which means it almost always gives you heads, then that would be a case of low variance. There's low variability in the results. There's little uncertainty about what's going to happen. It's going to be mostly heads with some occasional tails. So p equals $1/2$. Fair coin, that's the coin which is the most uncertain of all coins, in some sense. And it corresponds to the biggest variance. It corresponds to an X that has the widest distribution.

Now that we're on a roll and we can calculate such hugely complicated sums in simple ways, let us try to push our luck and do a problem with this flavor, but a little harder than that. So you go to one of those old-fashioned cocktail parties. All males at least will have those standard big hats which look identical. They check them in when they walk in. And when they walk out, since they look pretty identical, they just pick a random hat and go home. So n people, they pick their hats completely at random, quote, unquote, and then leave. And the question is, to say something about the number of people who end up, by accident or by luck, to get back their own hat, the exact same hat that they checked in.

OK, first what do we mean completely at random? Completely at random, we basically mean that any permutation of the hats is equally likely. Any way of distributing those n hats to the n people, any particular way is as likely as any other way. So there's complete symmetry between hats and people. So what we want to do is to calculate the expected value and the variance of this random variable X . Let's start with the expected value.

Let's reuse the trick from the binomial case. So total number of hats picked, we're going to think of total number of hats picked as a sum of $(0, 1)$ random variables. X_1 tells us whether person 1

got their own hat back. If they did, we record a 1. X_2 , the same thing. By adding all X 's is how many 1's did we get, which counts how many people selected their own hats. So we broke down the random variable of interest, the number of people who get their own hats back, as a sum of random variables. And these random variables, again, are easy to handle, because they're binary. The only take two values.

What's the probability that X_i is equal to 1, the i -th person has a probability that they get their own hat? There's n hats by symmetry. The chance is that they end up getting their own hat, as opposed to any one of the other $n - 1$ hats, is going to be $1/n$. So what's the expected value of X_i ? It's one times $1/n$. With probability $1/n$, you get your own hat, or you get a value of 0 with probability $1-1/n$, which is $1/n$.

All right, so we got the expected value of the X_i 's. And remember, we want to do is to calculate the expected value of X by using this decomposition? Are the random variables X_i independent of each other? You can try to answer that question by writing down a joint PMF for the X 's, but I'm sure that you will not succeed. But can you think intuitively? If I tell you information about some of the X_i 's, does it give you information about the remaining ones? Yeah. If I tell you that out of 10 people, 9 of them got their own hat back, does that tell you something about the 10th person? Yes. If 9 got their own hat, then the 10th must also have gotten their own hat back.

So the first 9 random variables tell you something about the 10th one. And conveying information of this sort, that's the case of dependence. All right, so the random variables are not independent. Are we stuck? Can we still calculate the expected value of X ? Yes, we can. And the reason we can is that expectations are linear. Expectation of a sum of random variables is the sum of the expectations. And that's always true. There's no independence assumption that's being used to apply that rule. So we have that the expected value of X is the sum of the expected value of the X_i 's. And this is a property that's always true. You don't need independence. You don't care. So we're adding n terms, each one of which has expected value $1/n$. And the final answer is 1.

So out of the 100 people who selected hats at random, on the average, you expect only one of them to end up getting their own hat back. Very good. So since we are succeeding so far, let's try to see if we can succeed in calculating the variance as well. And of course, we will. But it's going to be a little more complicated. The reason it's going to be a little more complicated is because the X_i 's are not independent, so the variance of the sum is not the same as the sum of the variances.

So it's not enough to find the variances of the X_i 's. We'll have to do more work. And here's what's involved. Let's start with the general formula for the variance, which, as I mentioned before, it's usually the simpler way to go about calculating variances. So we need to calculate the expected value for X -squared, and subtract from it the expectation squared. Well, we already found the expected value of X . It's equal to 1. So 1-squared gives us just 1. So we're left with the task of calculating the expected value of X -squared, the random variable X -squared. Let's try to follow the same idea. Write this messy random variable, X -squared, as a sum of hopefully simpler random variables.

So X is the sum of the X_i 's, so you square both sides of this. And then you expand the right-hand side. When you expand the right-hand side, you get the squares of the terms that appear here. And then you get all the cross-terms. For every pair of (i,j) that are different, i different than j , you're going to have a cross-term in the sum. So now, in order to calculate the expected value of X -squared, what does our task reduce to? It reduces to calculating the expected value of this term and calculating the expected value of that term. So let's do them one at a time.

Expected value of X_i squared, what is it going to be? Same trick as before. X_i takes value 0 or 1, so X_i squared takes just the same values, 0 or 1. So that's the easy one. That's the same as expected value of X_i , which we already know to be $1/n$. So this gives us a first contribution down here. The expected value of this term is going to be what? We have n terms in the summation. And each one of these terms has an expectation of $1/n$. So we did a piece of the puzzle. So now let's deal with the second piece of the puzzle.

Let's find the expected value of X_i times X_j . Now by symmetry, the expected value of X_i times X_j is going to be the same no matter what i and j you see. So let's just think about X_1 and X_2 and try to find the expected value of X_1 and X_2 . X_1 times X_2 is a random variable. What values does it take? Only 0 or 1? Since X_1 and X_2 are 0 or 1, their product can only take the values of 0 or 1. So to find the probability distribution of this random variable, it's just sufficient to find the probability that it takes the value of 1. Now, what does X_1 times X_2 equal to 1 mean? It means that X_1 was 1 and X_2 was 1. The only way that you can get a product of 1 is if both of them turned out to be 1's.

So that's the same as saying, persons 1 and 2 both picked their own hats. The probability that person 1 and person 2 both pick their own hats is the probability of two things happening, which is the product of the first thing happening times the conditional probability of the second, given that the first happened. And in words, this is the probability that the first person picked their own hat times the probability that the second person picks their own hat, given that the first person already picked their own. So what's the probability that the first person picks their own hat? We know that it's $1/n$.

Now, how about the second person? If I tell you that one person has their own hat, and that person takes their hat and goes away, from the point of view of the second person, there's $n - 1$ people left looking at $n - 1$ hats. And they're getting just hats at random. What's the chance that I will get my own? It's $1/(n - 1)$. So think of them as person 1 goes, picks a hat at random, it happens to be their own, and it leaves. You're left with $n - 1$ people, and there are $n - 1$ hats out there.

Person 2 goes and picks a hat at random, with probability $1/(n - 1)$, is going to pick his own hat. So the expected value now of this random variable is, again, that same number, because this is a 0, 1 random variable. So this is the same as expected value of X_i times X_j when i different than j . So here, all that's left to do is to add the expectations of these terms. Each one of these terms has an expected value that's $1/n$ times $(1/(n - 1))$.

And how many terms do we have? How many of these are we adding up? It's n -squared - n . When you expand the quadratic, there's a total of n -squared terms. Some are self-terms, n of them. And the remaining number of terms is n -squared - n . So here we got n -squared - n terms.

And so we need to multiply here with n -squared - n . And after you realize that this number here is 1, and you realize that this is the same as the denominator, you get the answer that the expected value of X squared equals 2. And then, finally going up to the top formula, we get the expected value of X squared, which is 2 - 1, and the variance is just equal to 1.

So the variance of this random variable, number of people who get their own hats back, is also equal to 1, equal to the mean. Looks like magic. Why is this the case? Well, there's a deeper explanation why these two numbers should come out to be the same. But this is something that would probably have to wait a couple of chapters before we could actually explain it. And so I'll stop here.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

LECTURE 7

- **Readings:** Finish Chapter 2

Lecture outline

- Multiple random variables
 - Joint PMF
 - Conditioning
 - Independence
- More on expectations
- Binomial distribution revisited
- A hat problem

Review

$$p_X(x) = \mathbf{P}(X = x)$$

$$p_{X,Y}(x,y) = \mathbf{P}(X = x, Y = y)$$

$$p_{X|Y}(x | y) = \mathbf{P}(X = x | Y = y)$$

$$p_X(x) = \sum_y p_{X,Y}(x,y)$$

$$p_{X,Y}(x,y) = p_X(x)p_{Y|X}(y | x)$$

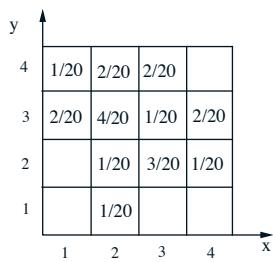
Independent random variables

$$p_{X,Y,Z}(x,y,z) = p_X(x)p_{Y|X}(y | x)p_{Z|X,Y}(z | x,y)$$

- Random variables X, Y, Z are independent if:

$$p_{X,Y,Z}(x,y,z) = p_X(x) \cdot p_Y(y) \cdot p_Z(z)$$

for all x, y, z



- Independent?
- What if we condition on $X \leq 2$ and $Y \geq 3$?

Expectations

$$\mathbf{E}[X] = \sum_x x p_X(x)$$

$$\mathbf{E}[g(X, Y)] = \sum_x \sum_y g(x, y) p_{X,Y}(x, y)$$

- In general: $\mathbf{E}[g(X, Y)] \neq g(\mathbf{E}[X], \mathbf{E}[Y])$
- $\mathbf{E}[\alpha X + \beta] = \alpha \mathbf{E}[X] + \beta$
- $\mathbf{E}[X + Y + Z] = \mathbf{E}[X] + \mathbf{E}[Y] + \mathbf{E}[Z]$
- If X, Y are independent:
 - $\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y]$
 - $\mathbf{E}[g(X)h(Y)] = \mathbf{E}[g(X)] \cdot \mathbf{E}[h(Y)]$

Variances

- $\text{Var}(aX) = a^2\text{Var}(X)$

- $\text{Var}(X + a) = \text{Var}(X)$

- Let $Z = X + Y$.

If X, Y are independent:

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

- Examples:

- If $X = Y$, $\text{Var}(X + Y) =$

- If $X = -Y$, $\text{Var}(X + Y) =$

- If X, Y indep., and $Z = X - 3Y$,
 $\text{Var}(Z) =$

Binomial mean and variance

- $X = \#$ of successes in n independent trials

- probability of success p

$$E[X] = \sum_{k=0}^n k \binom{n}{k} p^k (1-p)^{n-k}$$

- $X_i = \begin{cases} 1, & \text{if success in trial } i, \\ 0, & \text{otherwise} \end{cases}$

- $E[X_i] =$

- $E[X] =$

- $\text{Var}(X_i) =$

- $\text{Var}(X) =$

The hat problem

- n people throw their hats in a box and then pick one at random.

- X : number of people who get their own hat

- Find $E[X]$

$$X_i = \begin{cases} 1, & \text{if } i \text{ selects own hat} \\ 0, & \text{otherwise.} \end{cases}$$

- $X = X_1 + X_2 + \dots + X_n$

- $P(X_i = 1) =$

- $E[X_i] =$

- Are the X_i independent?

- $E[X] =$

Variance in the hat problem

- $\text{Var}(X) = E[X^2] - (E[X])^2 = E[X^2] - 1$

$$X^2 = \sum_i X_i^2 + \sum_{i,j:i \neq j} X_i X_j$$

- $E[X_i^2] =$

$$P(X_1 X_2 = 1) = P(X_1 = 1) \cdot P(X_2 = 1 | X_1 = 1)$$

$$=$$

- $E[X^2] =$

- $\text{Var}(X) =$

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Recitation 7
September 30, 2010

1. **Problem 2.35, page 130 in the text.** Verify the expected value rule

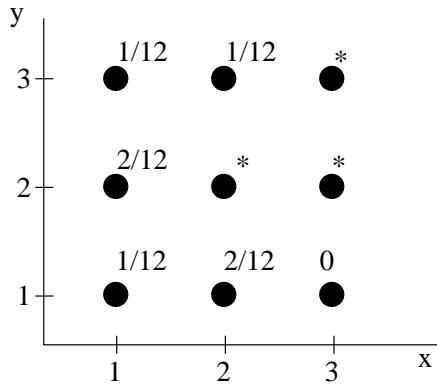
$$\mathbf{E}[g(X, Y)] = \sum_x \sum_y g(x, y) p_{X,Y}(x, y),$$

using the expected value rule for a function of a single random variable. Then, use the rule for the special case of a linear function, to verify the formula

$$\mathbf{E}[aX + bY] = a\mathbf{E}[X] + b\mathbf{E}[Y],$$

where a and b are given scalars.

2. Random variables X and Y can take any value in the set $\{1, 2, 3\}$. We are given the following information about their joint PMF, where the entries indicated by a * are left unspecified:



- (a) What is $p_X(1)$?
- (b) Provide a clearly labeled sketch of the conditional PMF of Y given that $X = 1$.
- (c) What is $\mathbf{E}[Y | X = 1]$?
- (d) Is there a choice for the unspecified entries that would make X and Y independent?

Let B be the event that $X \leq 2$ and $Y \leq 2$. We are told that conditioned on B , the random variables X and Y are independent.

- (e) What is $p_{X,Y}(2, 2)$?
 (If there is not enough information to determine the answer, say so.)
- (f) What is $p_{X,Y|B}(2, 2 | B)$?
 (If there is not enough information to determine the answer, say so.)

3. **Problem 2.33, page 128 in the text.** A coin that has probability of heads equal to p is tossed successively and independently until a head comes twice in a row or a tail comes twice in a row. Find the expected value of the number of tosses.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Recitation 7 Solutions
September 30, 2010

1. See the textbook, Problem 2.35, page 130.

2. (a)

$$\begin{aligned} p_X(1) &= \mathbf{P}(X = 1, Y = 1) + \mathbf{P}(X = 1, Y = 2) + \mathbf{P}(X = 1, Y = 3) \\ &= 1/12 + 2/12 + 1/12 = 1/3 \end{aligned}$$

(b) The solution is a sketch of the following conditional PMF:

$$p_{Y|X}(y | 1) = \frac{p_{Y,X}(y, 1)}{p_X(1)} = \begin{cases} 1/4, & \text{if } y = 1, \\ 1/2, & \text{if } y = 2, \\ 1/4, & \text{if } y = 3, \\ 0, & \text{otherwise.} \end{cases}$$

(c) $\mathbf{E}[Y | X = 1] = \sum_{y=1}^3 y p_{Y|X}(y | 1) = 1 \cdot \frac{1}{4} + 2 \cdot \frac{1}{2} + 3 \cdot \frac{1}{4} = 2$

(d) Assume that X and Y are independent. Because $p_{X,Y}(3, 1) = 0$ and $p_Y(1) = 1/4$, $p_X(3)$ must equal zero. This further implies $p_{X,Y}(3, 2) = 0$ and $p_{X,Y}(3, 3) = 0$. All the remaining probability mass must go to $(X, Y) = (2, 2)$, making $p_{X,Y}(2, 2) = 5/12$, $p_X(2) = 8/12$, and $p_Y(2) = 7/12$. However, $p_{X,Y}(2, 2) \neq p_X(2) \cdot p_Y(2)$, contradicting the assumption; thus X and Y are not independent.

A simpler explanation uses only two X values and two Y values for which all four (X, Y) pairs have specified probabilities. Note that if X and Y are independent, then $p_{X,Y}(1, 3)/p_{X,Y}(1, 1)$ and $p_{X,Y}(2, 3)/p_{X,Y}(2, 1)$ must be equal because they must both equal $p_Y(3)/p_Y(1)$. This necessary equality does not hold, so X and Y are not independent.

(e) Knowing that X and Y are conditionally independent given B , we must have

$$\frac{p_{X,Y}(1, 1)}{p_{X,Y}(1, 2)} = \frac{p_{X,Y}(2, 1)}{p_{X,Y}(2, 2)}$$

since the (X, Y) pairs in the equality are all in B . Thus

$$p_{X,Y}(2, 2) = \frac{p_{X,Y}(1, 2)p_{X,Y}(2, 1)}{p_{X,Y}(1, 1)} = \frac{(2/12)(2/12)}{1/12} = \frac{4}{12} = \frac{1}{3}.$$

(f) Since $\mathbf{P}(B) = 9/12 = 3/4$, we normalize to obtain $p_{X,Y|B}(2, 2) = \frac{p_{X,Y}(2, 2)}{\mathbf{P}(B)} = 4/9$.

3. See the textbook, Problem 2.33, page 128.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Tutorial 3
September 30/October 1, 2010

1. Let X and Y be independent random variables. Random variable X has mean μ_X and variance σ_X^2 , and random variable Y has mean μ_Y and variance σ_Y^2 . Let $Z = 2X - 3Y$. Find the mean and variance of Z in terms of the means and variances of X and Y .
2. Problem 2.40, page 133 in the text.
A particular professor is known for his arbitrary grading policies. Each paper receives a grade from the set $\{A, A-, B+, B, B-, C+\}$, with equal probability, independently of other papers. How many papers do you expect to hand in before you receive each possible grade at least once?
3. The joint PMF of the random variables X and Y is given by the following table:

$y = 3$	c	c	$2c$
$y = 2$	$2c$	0	$4c$
$y = 1$	$3c$	c	$6c$
	$x = 1$	$x = 2$	$x = 3$

- (a) Find the value of the constant c .
- (b) Find $p_Y(2)$.
- (c) Consider the random variable $Z = YX^2$. Find $\mathbf{E}[Z \mid Y = 2]$.
- (d) Conditioned on the event that $X \neq 2$, are X and Y independent? Give a one-line justification.
- (e) Find the conditional variance of Y given that $X = 2$.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Tutorial 3: Solutions

1. In general we have that $\mathbf{E}[aX + bY + c] = a\mathbf{E}[X] + b\mathbf{E}[Y] + c$. Therefore,

$$\mathbf{E}[Z] = 2 \cdot \mathbf{E}[X] - 3 \cdot \mathbf{E}[Y].$$

For the case of independent random variables, we have that if $Z = a \cdot X + b \cdot Y$, then

$$\text{var}(Z) = a^2 \cdot \text{var}(X) + b^2 \cdot \text{var}(Y).$$

Therefore, $\text{var}(Z) = 4 \cdot \text{var}(X) + 9 \cdot \text{var}(Y)$.

2. See online solutions.
3. (a) We can find c knowing that the probability of the entire sample space must equal 1.

$$\begin{aligned} 1 &= \sum_{x=1}^3 \sum_{y=1}^3 p_{X,Y}(x,y) \\ &= c + c + 2c + 2c + 4c + 3c + c + 6c \\ &= 20c \end{aligned}$$

Therefore, $c = \frac{1}{20}$.

$$(b) p_Y(2) = \sum_{x=1}^3 p_{X,Y}(x,2) = 2c + 0 + 4c = 6c = \frac{3}{10}.$$

$$(c) Z = YX^2$$

$$\begin{aligned} \mathbf{E}[Z \mid Y = 2] &= \mathbf{E}[YX^2 \mid Y = 2] \\ &= \mathbf{E}[2X^2 \mid Y = 2] \\ &= 2\mathbf{E}[X^2 \mid Y = 2] \end{aligned}$$

$$p_{X|Y}(x \mid 2) = \frac{p_{X,Y}(x,2)}{p_Y(2)}.$$

Therefore,

$$p_{X|Y}(x \mid 2) = \begin{cases} \frac{1/10}{3/10} = \frac{1}{3} & \text{if } x = 1 \\ \frac{1/5}{3/10} = \frac{2}{3} & \text{if } x = 3 \\ 0 & \text{otherwise} \end{cases}$$

$$\begin{aligned} \mathbf{E}[Z \mid Y = 2] &= 2 \sum_{x=1}^3 x^2 p_{X|Y}(x \mid 2) \\ &= 2 \left((1^2) \cdot \frac{1}{3} + (3^2) \cdot \frac{2}{3} \right) \\ &= \frac{38}{3} \end{aligned}$$

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
 (Fall 2010)

(d) Yes. Given $X \neq 2$, the distribution of X is the same given $Y = y$.

$$\mathbf{P}(X = x | Y = y, X \neq 2) = \mathbf{P}(X = x | X \neq 2).$$

For example,

$$\mathbf{P}(X = 1 | Y = 1, X \neq 2) = \mathbf{P}(X = 1 | Y = 3, X \neq 2) = \mathbf{P}(X = 1 | X \neq 2) = \frac{1}{3}$$

(e) $p_{Y|X}(y | 2) = \frac{p_{X,Y}(2,y)}{p_X(2)}$.

$$p_X(2) = \sum_{y=1}^3 p_{X,Y}(2,y) = c + 0 + c = 2c = \frac{1}{10}.$$

Therefore,

$$p_{Y|X}(y | 2) = \begin{cases} \frac{1/20}{1/10} = \frac{1}{2} & \text{if } y = 1 \\ \frac{1/20}{1/10} = \frac{1}{2} & \text{if } y = 3 \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbf{E}[Y^2 | X = 2] = \sum_{y=1}^3 y^2 p_{Y|X}(y | 2) = (1^2) \cdot \frac{1}{2} + (3^2) \cdot \frac{1}{2} = 5.$$

$$\mathbf{E}[Y | X = 2] = \sum_{y=1}^3 y p_{Y|X}(y | 2) = 1 \cdot \frac{1}{2} + 3 \cdot \frac{1}{2} = 2.$$

$$\text{var}(Y | X = 2) = \mathbf{E}[Y^2 | X = 2] - \mathbf{E}[Y | X = 2]^2 = 5 - 2^2 = 1.$$

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

$$\begin{aligned}
 \mathbf{P}(\text{ripe} \cap \text{gala}) &= \mathbf{P}(\text{ripe} \cap \text{gala} \mid \text{Caleb's ripe gala})\mathbf{P}(\text{Caleb's ripe gala}) \\
 &\quad + \mathbf{P}(\text{ripe} \cap \text{gala} \mid k \text{ gala})\mathbf{P}(k \text{ gala}) \\
 &\quad + \mathbf{P}(\text{ripe} \cap \text{gala} \mid n - k \text{ honey crisp})\mathbf{P}(n - k \text{ honey crisp}) \\
 &= 1 \cdot \frac{1}{n+1} + g \cdot \frac{k}{n+1} + 0 \cdot \frac{n-k}{n+1} \\
 &= \frac{1+gk}{n+1}.
 \end{aligned}$$

Combining this result with that of (i),

$$\mathbf{P}(\text{gala} \mid \text{ripe apple}) = \frac{\frac{1+gk}{n+1}}{\frac{1+gk+h(n-k)}{n+1}} = \frac{1+gk}{1+gk+h(n-k)}.$$

- (c) **(10 points)** Let A be the event that the first 10 apples picked were all gala and let B be the event that exactly 10 gala apples were picked out of the 20 apples.

$$\mathbf{P}(A \mid B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)}.$$

The probability of the event B can be computed by using the result in (a), where $n = 20$ and $k = 10$. $\mathbf{P}(B) = \binom{20}{10}p^{10}(1-p)^{10}$. $\mathbf{P}(A \cap B)$ can be computed by separating the event of A and B into disjoint events. The event $\{A \cap B\} = \{\text{1st 10 apples picked are gala and the last 10 apples are honey crisp}\}$. $\mathbf{P}(A \cap B) = p^{10}(1-p)^{10}$. Therefore,

$$\mathbf{P}(A \mid B) = \frac{p^{10}(1-p)^{10}}{\binom{20}{10}p^{10}(1-p)^{10}} = \frac{1}{\binom{20}{10}}.$$

3. (a) **(10 points)** Bob picks apples from each tree until he finds one that is not ripe, and so if we think of the event “picking an unripe apple” as a “success”, then X_i is the number of apples picked (i.e. trials) until we get our first success. Therefore X_i is a geometrically distributed random variable with probability of success $1-g$.

Since each tree has a random collection of apples, the apples Bob picks from the first tree are independent of the apples he picks from the second tree, and therefore X_1 and X_2 are independent and identically distributed random variables.

The PMF of X_i is

$$\begin{aligned}
 p_{X_1}(k) = p_{X_2}(k) &= (1 - (1-g))^{k-1}(1-g) \\
 &= \begin{cases} g^{k-1}(1-g), & k = 1, 2, 3, \dots, \\ 0, & \text{otherwise,} \end{cases}
 \end{aligned}$$

the expectation of X_i is

$$\mathbf{E}[X_1] = \mathbf{E}[X_2] = \frac{1}{1-g},$$

and the variance of X_i is

$$\begin{aligned}
 \text{var}(X_1) = \text{var}(X_2) &= \frac{1 - (1-g)}{(1-g)^2} \\
 &= \frac{g}{(1-g)^2}.
 \end{aligned}$$

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz 1 Solutions | Fall 2009)

- (b) **(12 points)** Given that $Y_2 = (X_1 - 1) + (X_2 - 1)$, we can use the linearity property of expectations to find the expectation of Y_2 .

$$\begin{aligned}\mathbf{E}[Y_2] &= \mathbf{E}[(X_1 - 1) + (X_2 - 1)] \\ &= \mathbf{E}[X_1] + \mathbf{E}[X_2] - 2 \\ &= \frac{2}{1-g} - 2 \\ &= \frac{2g}{1-g}.\end{aligned}$$

Since adding or subtracting a constant from a random variable has no effect on its variance, and since X_1 and X_2 are independent, we have

$$\begin{aligned}\text{var}(Y_2) &= \text{var}((X_1 - 1) + (X_2 - 1)) \\ &= \text{var}(X_1 + X_2) \\ &= \text{var}(X_1) + \text{var}(X_2) \\ &= \frac{2g}{(1-g)^2}.\end{aligned}$$

- (c) **(12 points)**

Let $k \geq 0$ and $\ell \geq k$. The event “ $Y_1 = k$ and $Y_2 = \ell$ ” is identical to the event “ $X_1 = k+1$ and $X_2 = \ell-k+1$ ”. Since X_1 and X_2 are independent, the desired probability is $(1-g)g^{k+1-1}(1-g)g^{\ell-k+1-1} = (1-g)^2g^\ell$, when $0 \leq k \leq \ell$, and zero otherwise.

$$p_{Y_1, Y_2}(k, \ell) = \begin{cases} g^\ell(1-g)^2, & k = 0, 1, \dots, \ell, \text{ and } \ell = k, k+1, \dots, \\ 0, & \text{otherwise.} \end{cases}$$

Note that we can interpret this result as the probability of ℓ failures (i.e. picking ℓ ripe apples) and two successes (i.e. the two unripe apples that cause Bob to stop at each tree). Note also that although the expression for the joint PMF doesn't depend on k , the sample space does depend on k because $\ell \geq k$: Bob cannot pick more ripe apples in the first tree than he does in the first two trees combined.

- (d) (i) **(5 points)** No. X_1 and Y_2 are not independent because $Y_2 \geq X_1 - 1$ and so knowing X_1 gives us information about Y_2 . Specifically, if we know that Bob picked 10 total apples from the first tree (i.e. $X_1 = 10$), then we also know that Bob must have picked at least 9 ripe apples from the first two trees combined (i.e. $Y_2 \geq 9$).
(ii) **(5 points)** Yes. X_2 relates only to the second tree and Y_1 relates only to the first tree, and since the picking of apples is independent across trees, the two random variables are independent.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Question 1

Multiple Choice Questions: **CLEARLY** circle the appropriate choice. Scratch paper is available if needed, though **NO** partial credit will be given for the Multiple Choice. **Each multiple choice question is worth 4 points.**

a. Which of the following statements is NOT true?

- (i) If $A \subset B$, then $\mathbf{P}(A) \leq \mathbf{P}(B)$.
- (ii) $\mathbf{P}(B) > 0$, then $\mathbf{P}(A|B) \geq \mathbf{P}(A)$.
- (iii) $\mathbf{P}(A \cap B) \geq \mathbf{P}(A) + \mathbf{P}(B) - 1$.
- (iv) $\mathbf{P}(A \cap B^c) = \mathbf{P}(A \cup B) - \mathbf{P}(B)$.

Solution: A counterexample: if we have two events A, B such that $P(B) > 0$ and $P(A) > 0$, but $A \cap B = \emptyset$, then $P(A|B) = 0$, but $P(A) > P(A|B)$. It's easy to come up with examples like this: for example, take any sample space with event A such that $P(A) > 0$, and $P(A^c) > 0$, it follows that $P(A|A^c) = 0$, but $P(A) > 0$.

b. We throw n identical balls into m urns at random, where each urn is equally likely and each throw is independent of any other throw. What is the probability that the i -th urn is empty?

- (i) $\left(1 - \frac{1}{m}\right)^n$
- (ii) $\left(1 - \frac{1}{n}\right)^m$
- (iii) $\binom{m}{n} \left(1 - \frac{1}{n}\right)^m$
- (iv) $\binom{n}{m} \left(\frac{1}{m}\right)^n$

Solution: The probability of the j th ball going into the i th urn is $1/m$. Hence, the probability of the j th ball not going into the i th urn is $(1 - 1/m)$. Since all throws are independent from one another, we can multiply these probabilities: the probability of all n balls not going into the i th urn, i.e. it is empty, is $(1 - \frac{1}{m})^n$.

c. We toss two fair coins simultaneously and independently. If the outcomes of the two coin tosses are the same, we win; otherwise, we lose. Let A be the event that the first coin comes up heads, B be the event that the second coin comes up heads, and C be the event that we win. Which of the following statements is true?

- (i) Events A and B are not independent.
- (ii) A and C are independent.
- (iii) Events A and B are conditionally independent given C .
- (iv) The probability of winning is $3/4$.

Solution: The sample space in this case is $\Omega = \{(H, H), (H, T), (T, H), (T, T)\}$. The probability law is a uniform distribution over this space. We have $A = \{(H, H), (H, T)\}$, $B = \{(H, H), (T, H)\}$, and $C = \{(H, H), (T, T)\}$. By the discrete uniform law, $P(A) = P(B) =$

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz 1 Solutions | Spring 2009)

$P(C) = 1/2$. We also have $P(A \cap C) = 1/4$, hence $P(A \cap C) = P(A)P(C)$, and the two events are independent. Intuitively, knowing that you won adds no information about whether your coin turned up heads or not: stating this formally, we have $P(A|C) = P(A)$.

- d. For a biased coin, the probability of “heads” is $1/3$. Let H be the number of heads in five independent coin tosses. What is the probability $\mathbf{P}(\text{first toss is a head} \mid H = 1 \text{ or } H = 5)$?

(i) $\frac{\frac{1}{3}(\frac{2}{3})^4}{5\frac{1}{3}(\frac{2}{3})^4 + (\frac{1}{3})^5}$

(ii) $\frac{\frac{1}{3}(\frac{2}{3})^4}{\frac{1}{3}(\frac{2}{3})^4 + (\frac{1}{3})^5}$

(iii) $\boxed{\frac{\frac{1}{3}(\frac{2}{3})^4 + (\frac{1}{3})^5}{5\frac{1}{3}(\frac{2}{3})^4 + (\frac{1}{3})^5}}$

(iv) $\frac{1}{5}$

Solution: Let A be the event that the first toss is a head.

$$\begin{aligned} P(A|\{H = 1\} \text{ or } \{H = 5\}) &= \frac{P(A \cap (\{H = 1\} \cup \{H = 5\}))}{P(\{H = 1\} \cup \{H = 5\})} \\ &= \frac{P((A \cap \{H = 1\}) \cup (A \cap \{H = 5\}))}{P(\{H = 1\} \cup \{H = 5\})} \\ &= \frac{P(\{H = 5\}) + P(A \cap \{H = 1\})}{P(\{H = 1\}) + P(\{H = 5\})} \\ &= \frac{(1/3)^5 + (1/3)(2/3)^4}{\binom{5}{1}(1/3)(2/3)^4 + \binom{5}{5}(1/3)^5}. \end{aligned}$$

- e. A well-shuffled deck of 52 cards is dealt evenly to two players (26 cards each). What is the probability that player 1 gets all the aces?

(i) $\boxed{\frac{\binom{48}{22}}{\binom{52}{26}} = \frac{26 \times 25 \times 24 \times 23}{52 \times 51 \times 50 \times 49}}$

(ii) $\frac{4 \binom{48}{22}}{\binom{52}{26}} = 4 \times \frac{26 \times 25 \times 24 \times 23}{52 \times 51 \times 50 \times 49}$

(iii) $\frac{48!}{22!} \frac{52!}{26!}$

(iv) $\frac{4! \binom{48}{22}}{\binom{52}{26}} = 4! \times \frac{26 \times 25 \times 24 \times 23}{52 \times 51 \times 50 \times 49}$

Solution: Let A be the event that player 1 gets all aces. By the discrete uniform law,

$$P(A) = \frac{|A|}{|\Omega|}. \quad (1)$$

$|\Omega| = \binom{52}{26}$ is the number of hands (26 cards from 52) player 1 can have. Additionally, once we have given player 1 all aces, then they must be given an additional 22 cards from the remaining 48 cards in the deck. Hence,

$$P(A) = \frac{\binom{48}{22}}{\binom{52}{26}}$$

f. Suppose X, Y and Z are three independent discrete random variables. Then, X and $Y + Z$ are

- (i) always
- (ii) sometimes
- (iii) never

independent.

Solution: Since X is independent of Y and Z , X is independent of $g(Y, Z)$ for any function $g(Y, Z)$, including $g(Y, Z) = Y + Z$ (see page 114 of the book).

g. To obtain a driving licence, Mina needs to pass her driving test. Every time Mina takes a driving test, with probability $1/2$, she will clear the test independent of her past. Mina failed her first test. Given this, let Y be the additional number of tests Mina takes before obtaining a licence. Then,

- (i) $E[Y] = 1$.
- (ii) $E[Y] = 2$.
- (iii) $E[Y] = 0$.

Solution: Y is defined as the number of additional tests Mina takes, so this is independent of the fact that she failed her first test. Y is a geometric RV with $p = 1/2$. Hence, $E[Y] = 1/p = 2$.

h. Consider two random variables X and Y , each taking values in $\{1, 2, 3\}$. Let their joint PMF be such that for any $1 \leq x, y \leq 3$, $P_{X,Y}(x, y) = 0$ if $(x, y) \in \{(1, 3), (2, 1), (3, 2)\}$, and $P_{X,Y}(x, y) > 0$ if $(x, y) \in \{(1, 1), (1, 2), (2, 2), (2, 3), (3, 1), (3, 3)\}$. Then,

- (i) X and Y can be independent or dependent depending upon the values of $P_{X,Y}(x, y)$ for $(x, y) \in \{(1, 1), (1, 2), (2, 2), (2, 3), (3, 1), (3, 3)\}$.
- (ii) X and Y are always independent.
- (iii) X and Y can never be independent.

Solution: If, for example, we are given information that $X = 1$, we know that Y can never take value 3. However, without this information about X the probability $p_Y(3)$ is strictly positive and so $p_{Y|X}(y|x) \neq p_Y(y)$, for $x = 1$ and $y = 3$, i.e. X and Y can never be independent.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz 1 Solutions | Spring 2009)

i. Suppose you play a *matching coins* game with your friend as follows. Both you and your friend each have your own coin. Each time, the two of you reveal a side (i.e. H or T) of your coin to each other simultaneously. If the sides match, you WIN \$1 from your friend and if sides do not match then you lose \$1 to your friend. Your friend has a complicated (unknown) strategy in selecting the sides over time. You decide to go with the following simple strategy. Every time, you will toss your unbiased coin independently of everything else, and you will reveal its outcome to your friend (of course, your friend does not know the outcome of your random toss until you reveal it). Then,

- (i) On average, you will lose money to your smart friend.
- (ii) On average, you will neither lose nor win. That is, your average gain/loss is 0.
- (iii) On average, you will make money from your friend.

Solution: Let X_i be a random variable denoting your winnings at the i 'th round of the game, i.e., $X_i = 1$ if you win, $X_i = -1$ if you lose. At each round your friend chooses either heads or tails, using some strategy that you don't know about. The key property is that *for any choice that your friend makes, we have $p_{X_i}(1) = p_{X_i}(-1) = 0.5$* : i.e., we always have a 0.5 probability that our coin toss will match the choice made by our friend. It can be verified that $\mathbf{E}[X_i] = 0$, and hence your average gain/loss is 0.

j. Let $X_i, 1 \leq i \leq 4$ be independent Bernoulli random variables each with mean $p = 0.1$. Let $X = \sum_{i=1}^4 X_i$. Then,

- (i) $E[X_1|X = 2] = 0.1$.
- (ii) $E[X_1|X = 2] = 0.5$.
- (iii) $E[X_1|X = 2] = 0.25$.

Solution: We have $P(X_1 = 1|X = 2) = 0.5$, because

$$\begin{aligned} P(X_1 = 1|X = 2) &= \frac{P(X_1 = 1 \cap X = 2)}{P(X = 2)} \\ &= \frac{p \times \binom{3}{1}p(1-p)^2}{\binom{4}{2}p^2(1-p)^2} \\ &= \frac{\binom{3}{1}}{\binom{4}{2}} = 0.5 \end{aligned}$$

(Note that $\binom{4}{2}p^2(1-p)^2$ is the probability of seeing 2 heads out of 4 tosses, and $\binom{3}{1}p(1-p)^2$ is the probability of seeing 1 head in the last 3 tosses.)

Hence,

$$\mathbf{E}[X_1|X = 2] = 1 \times P(X_1 = 1|X = 2) + 0 \times P(X_1 = 0|X = 2) = 0.5$$

Question 2:

Alice and Bob both need to buy a bicycle. The bike store has a stock of four green, three yellow, and two red bikes. Alice randomly picks one of the bikes and buys it. Immediately after, Bob does the same. The sale price of the green, yellow, and red bikes are \$300, \$200 and \$100, respectively.

Let A be the event that Alice bought a green bike, and B be the event that Bob bought a green bike.

- a. (5 points) What is $\mathbf{P}(A)$? What is $\mathbf{P}(A|B)$?

Solution: We have $\mathbf{P}(A) = 4/9$ (4 green bikes out of 9), and $\mathbf{P}(A|B) = 3/8$ (since we know that Bob has a green bike, Alice can have one of 3 green bikes out of the remaining 8).

- b. (2 points) Are A and B independent events? Justify your answer.

Solution: Since $\mathbf{P}(A) \neq \mathbf{P}(A|B)$, the events are *not* independent. Informally, since there is a fixed quantity of green bikes, if Alice buys one, then the chances that Bob buys one too are slightly decreased.

- c. (5 points) What is the probability that at least one of them bought a green bike?

Solution: The requested probability is $\mathbf{P}(A \cup B)$. We have

$$\begin{aligned}\mathbf{P}(A \cup B) &= \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B) \\ &= \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A|B) \cdot \mathbf{P}(B) \\ &= \frac{4}{9} + \frac{4}{9} - \frac{3}{8} \cdot \frac{4}{9} = \frac{13}{18} = 0.722.\end{aligned}$$

d. (5 points) What is the probability that Alice and Bob bought bicycles of different colors?

Solution: Let's compute first the probability that Alice and Bob bought bikes of the same color. We have

$$\mathbf{P}(\{G, G\}) = \frac{4}{9} \cdot \frac{3}{8}, \quad \mathbf{P}(\{Y, Y\}) = \frac{3}{9} \cdot \frac{2}{8}, \quad \mathbf{P}(\{R, R\}) = \frac{2}{9} \cdot \frac{1}{8}.$$

Therefore, the probability of buying bikes of different color is

$$\mathbf{P}(\text{different color}) = 1 - \mathbf{P}(\text{same color}) = 1 - \left(\frac{12}{72} + \frac{6}{72} + \frac{2}{72} \right) = \frac{13}{18} = 0.722.$$

e. (5 points) Given that Bob bought a green bike, what is the expected value of the amount of money spent by Alice?

Solution: If Bob bought a green bike, then the conditional probabilities of Alice buying a green, yellow, or red bike are $\frac{3}{8}$, $\frac{3}{8}$ and $\frac{2}{8}$, respectively. The expected amount of money spent by Alice is therefore

$$\$300 \cdot \frac{3}{8} + \$200 \cdot \frac{3}{8} + \$100 \cdot \frac{2}{8} = \$212.50.$$

f. (5 points) Let G be the number of green bikes that remain in the store after Alice and Bob's visit. Compute $\mathbf{P}(B|G = 3)$.

Solution: If $G = 3$, then exactly one green bike was bought. By symmetry, there is equal chance that Alice or Bob bought it, thus $\mathbf{P}(B|G = 3) = \frac{1}{2}$. Alternatively, define $A \setminus B$ as the elements of A that are not in B . We have:

$$\begin{aligned} \mathbf{P}(B|G = 3) &= \mathbf{P}(B|\{A \setminus B\} \cup \{B \setminus A\}) \\ &= \frac{\mathbf{P}(B \setminus A)}{\mathbf{P}(\{A \setminus B\} \cup \{B \setminus A\})} = \frac{20/72}{40/72} = \frac{1}{2}. \end{aligned}$$

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz 1 Solutions | Spring 2009)

Question 3:

Magic Games Inc. is a store that sells all sorts of fun games. One of its popular products is its magic 4-sided dice. The dice come in pairs; each die can be fair or crooked, and the dice in any pair can function independently or, in some cases, can have magnets inside them that cause them to behave in unpredictable ways when rolled together.

Xavier and Yvonne together buy a pair of dice from this store. Each of them picks a die in the pair; one of them then rolls the two dice together. Let X be the outcome of Xavier's die and Y the outcome of Yvonne's die. The joint PMF of X and Y , $p_{X,Y}(x,y)$, is given by the following figure:

	4	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{2}{20}$
	3	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{2}{20}$	$\frac{1}{20}$
Y	2	$\frac{1}{20}$	$\frac{2}{20}$	$\frac{1}{20}$	$\frac{1}{20}$
	1	$\frac{2}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$
		1	2	3	4
		X			

- (a) (2 points) Find the PMF of the outcome of Xavier's die, $p_X(x)$.

Solution:

$$p_X(x) = \sum_{y=1}^4 p_{X,Y}(x,y) = \begin{cases} \frac{1}{4} & x = 1, 2, 3, 4 \\ 0 & o.w. \end{cases}$$

- (b) (2 points) Find the PMF of the outcome of Yvonne's die, $p_Y(y)$.

Solution:

$$p_Y(y) = \sum_{x=1}^4 p_{X,Y}(x,y) = \begin{cases} \frac{1}{4} & y = 1, 2, 3, 4 \\ 0 & o.w. \end{cases}$$

- (c) (2 points) Are X and Y independent?

Solution: No. One of many counter examples: $p_X(x)$ does not equal $p_{X|Y}(x|2)$.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz 1 Solutions | Spring 2009)

Zach and Wendy are intrigued by Xavier and Yvonne's dice. They visit the store and buy a pair of dice of their own. Again, each of them picks a die in the pair; one of them then rolls the two dice together. Let Z be the outcome of Zach's die and W the outcome of Wendy's die. The joint PMF of Z and W , $p_{Z,W}(z,w)$, is given by the following figure:

	4	$\frac{2}{24}$	$\frac{1}{24}$	$\frac{2}{24}$	$\frac{1}{24}$
	3	$\frac{2}{24}$	$\frac{1}{24}$	$\frac{2}{24}$	$\frac{1}{24}$
W	2	$\frac{2}{24}$	$\frac{1}{24}$	$\frac{2}{24}$	$\frac{1}{24}$
	1	$\frac{2}{24}$	$\frac{1}{24}$	$\frac{2}{24}$	$\frac{1}{24}$
		1	2	3	4
			Z		

The store also sells a variety of magic coins, some fair and some crooked. Alice buys a coin that on each toss comes up heads with probability $3/4$.

- (d) (5 points) Wondering whether to buy some dice as well, Alice decides to try out her friends' dice first. She does the following. First, she tosses her coin. If the coin comes up heads, she borrows Xavier and Yvonne's dice pair and rolls the two dice. If the coin comes up tails, she borrows Zach and Wendy's dice pair and rolls those instead. What is the probability that she rolls a double, i.e., that both dice in the pair she rolls show the same number?

Solution: Let event D be the set of all doubles, and let event A be the event that Alice's coin toss results in heads. Using the law of total probability:

$$\begin{aligned}
 P(D) &= P(D|A)P(A) + P(D|A^c)P(A^c) \\
 &= \frac{3}{4} \times \frac{4}{10} + \frac{1}{4} \times \frac{1}{4} \\
 &= \frac{58}{160} = .3625
 \end{aligned}$$

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz 1 Solutions | Spring 2009)

- (e) (5 points) Alice has still not made up her mind about the dice. She tries another experiment. First, she tosses her coin. If the coin comes up heads, she takes Xavier and Yvonne's dice pair and rolls the dice repeatedly until she gets a double; if the coin comes up tails, she does the same with Zach and Wendy's dice. What is the expected number of times she will need to roll the dice pair she chooses? (Assume that if a given pair of dice is rolled repeatedly, the outcomes of the different rolls are independent.)

Solution: Let random variable N be the number of rolls until doubles is rolled. The distribution on N condition on the set of dice being rolled is a geometric random variable. Using the total expectation theorem, the expected value of N is:

$$\begin{aligned} E[N] &= E[N|A]P(A) + E[N|A^c]P(A^c) \\ &= \frac{3}{4} \times \frac{1}{\frac{4}{10}} + \frac{1}{4} \times \frac{1}{\frac{1}{4}} \\ &= \frac{23}{8} = 2.875 \end{aligned}$$

- (f) (5 points) Alice is bored with the dice and decides to experiment with her coin instead. She tosses the coin until she has seen a total of 11 heads. Let R be the number of tails she sees. Find $E[R]$. (Assume independent tosses.)

Solution: The time T until Alice sees a total of 11 heads is the sum of 11 independent and identically distributed geometric random variables with parameter $p = \frac{3}{4}$. Random variable R , the number of tails she sees, is $T - 11$. Thus:

$$\begin{aligned} E[R] &= E[T] - 11 \\ &= 11 \times \frac{1}{\frac{3}{4}} - 11 \\ &= \frac{11}{12} = .9167 \end{aligned}$$

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz 1 Solutions | Spring 2009)

- (g) (5 points) Alice tries another experiment with her coin. Let A be the event that the second head she sees occurs on the 7th coin toss, and let S be the position of the first head. Find the conditional PMF of S given the event A , $p_{S|A}(s)$.

Solution: The probability of event A can be found by choosing one of the first 6 outcomes to be a head, the others tails, and then the outcome of the 7th toss to be head, which is $\binom{6}{1}(1-p)^5p^2$, where $p = \frac{3}{4}$. The intersection of $S = s$ with event A , $P(S = s \cap A)$, is an event with probability $(1-p)^5p^2$ for all values of s ($s = 1, \dots, 6$). Consequently, $p_{S|A}(s) = \frac{P(S=s \cap A)}{P(A)}$ is a uniform distribution over the range of s ($s = 1, \dots, 6$).

$$p_{S|A}(s) = \begin{cases} \frac{1}{6} & s = 1, \dots, 6 \\ 0 & \text{o.w.} \end{cases}$$

- (h) (5 points) Alice's friend Bob buys a coin from the same store that turns out to be fair, i.e., that on any toss comes up heads with probability $1/2$. He tosses the coin repeatedly until he has seen either a total of 11 heads or a total of 11 tails. Let U be the number of times he will need to toss the coin. Find the PMF of U , $p_U(u)$. (Assume independent tosses.)

Solution: Bob must toss a coin at least 11 times and at most 21 times in order to have either 11 heads or 11 tails. The intersection of Bob requiring u tosses and 11 of those tosses being heads, is the sum of probability the $\binom{u-1}{10}$ sequences that conclude with a head and have a total of 11 heads. The probability of each of those sequences is $(\frac{1}{2})^u$. If we consider any sequence in Bob's experiment with u tosses, since the coin is fair, that sequence is equally likely to have 11 heads and $u - 11$ tails or 11 tails and $u - 11$ heads. Consequently, the intersection of Bob requiring u tosses and 11 of those tosses being tails is identical to the probability that the sequence had 11 heads. Summing these two mutually exclusive probabilities which total $p_U(u)$:

$$p_U(u) = \begin{cases} 2\binom{u-1}{10}(\frac{1}{2})^u & u = 11, \dots, 21 \\ 0 & \text{o.w.} \end{cases}$$

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Summary of Results for Special Random Variables

Discrete Uniform over $[a, b]$:

$$p_X(k) = \begin{cases} \frac{1}{b-a+1}, & \text{if } k = a, a+1, \dots, b, \\ 0, & \text{otherwise,} \end{cases}$$

$$\mathbf{E}[X] = \frac{a+b}{2}, \quad \text{var}(X) = \frac{(b-a)(b-a+2)}{12}.$$

Bernoulli with Parameter p : (Describes the success or failure in a single trial.)

$$p_X(k) = \begin{cases} p, & \text{if } k = 1, \\ 1-p, & \text{if } k = 0, \end{cases}$$

$$\mathbf{E}[X] = p, \quad \text{var}(X) = p(1-p).$$

Binomial with Parameters p and n : (Describes the number of successes in n independent Bernoulli trials.)

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \dots, n,$$

$$\mathbf{E}[X] = np, \quad \text{var}(X) = np(1-p).$$

Geometric with Parameter p : (Describes the number of trials until the first success, in a sequence of independent Bernoulli trials.)

$$p_X(k) = (1-p)^{k-1} p, \quad k = 1, 2, \dots,$$

$$\mathbf{E}[X] = \frac{1}{p}, \quad \text{var}(X) = \frac{1-p}{p^2}.$$

Problem 1: (75 points)

Note: All parts can be done independently, with the exception of the last part. Just in case you made a mistake in the previous part, you can use a symbol for the expression you found there, and use that symbol in the formulas for the last part.

Note: Algebraic or numerical expressions do not need to be simplified in your answers.

Jon and Stephen cannot help but think about their commutes using probabilistic modeling. Both of them start promptly at 8am.

Stephen drives and thus is at the mercy of traffic lights. When all traffic lights on his route are green, the entire trip takes 18 minutes. Stephen's route includes 5 traffic lights, each of which is red with probability $1/3$, independent of every other light. Each red traffic light that he encounters adds 1 minute to his commute (for slowing, stopping, and returning to speed).

1. **(10 points)** Find the PMF, expectation, and variance of the length (in minutes) of Stephen's commute.
2. **(10 points)** Given that Stephen's commute took him at most 19 minutes, what is the expected number of red lights that he encountered?
3. **(10 points)** Given that the last red light encountered by Stephen was the fourth light, what is the conditional variance of the total number of red lights he encountered?
4. **(10 points)** Given that Stephen encountered a total of three red lights, what is the probability that exactly two out of the first three lights were red?

Jon's commuting behavior is rather simple to model. Jon walks a total of 20 minutes from his home to a station and from a station to his office. He also waits for X minutes for a subway train, where X has the discrete uniform distribution on $\{0, 1, 2, 3\}$. (All four values are equally likely, and independent of the traffic lights encountered by Stephen.)

5. **(5 points)** What is the PMF of the length of Jon's commute in minutes?
6. **(10 points)** Given that there was exactly one person arriving at **exactly** 8:20am, what is the probability that this person was Jon?
7. **(10 points)** What is the probability that Stephen's commute takes at most as long as Jon's commute?
8. **(10 points)** Given that Stephen's commute took at most as long as Jon's, what is the conditional probability that Jon waited 3 minutes for his train?

Problem 2. (30 points) For each one of the statements below, give either a proof or a counterexample showing that the statement is not always true.

1. **(10 points)** If events A and B are independent, then the events A and B^c are also independent.
2. **(10 points)** Let A , B , and C be events associated with a common probabilistic model, and assume that $0 < \mathbf{P}(C) < 1$. Suppose that A and B are conditionally independent given C . Then, A and B are conditionally independent given C^c .

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz 1 | Fall 2010)

3. **(10 points)** Let X and Y be independent random variables. Then, $\text{var}(X + Y) \geq \text{var}(X)$.

Each question is repeated in the following pages. Please write your answer on the appropriate page.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz 1 | Fall 2010)

Problem 1: (75 points)

Note: All parts can be done independently, with the exception of the last part. Just in case you made a mistake in the previous part, you can use a symbol for the expression you found there, and use that symbol in the formulas for the last part.

Note: Algebraic or numerical expressions do not need to be simplified in your answers.

Jon and Stephen cannot help but think about their commutes using probabilistic modeling. Both of them start promptly at 8am.

Stephen drives and thus is at the mercy of traffic lights. When all traffic lights on his route are green, the entire trip takes 18 minutes. Stephen's route includes 5 traffic lights, each of which is red with probability $1/3$, independent of every other light. Each red traffic light that he encounters adds 1 minute to his commute (for slowing, stopping, and returning to speed).

1. **(10 points)** Find the PMF, expectation, and variance of the length (in minutes) of Stephen's commute.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz 1 | Fall 2010)

2. **(10 points)** Given that Stephen's commute took him at most 19 minutes, what is the expected number of red lights that he encountered?

3. **(10 points)** Given that the last red light encountered by Stephen was the fourth light, what is the conditional variance of the total number of red lights he encountered?

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz 1 | Fall 2010)

4. **(10 points)** Given that Stephen encountered a total of three red lights, what is the probability that exactly two out of the first three lights were red?

Jon's commuting behavior is rather simple to model. Jon walks a total of 20 minutes from his home to a station and from a station to his office. He also waits for X minutes for a subway train, where X has the discrete uniform distribution on $\{0, 1, 2, 3\}$. (All four values are equally likely, and independent of the traffic lights encountered by Stephen.)

5. **(5 points)** What is the PMF of the length of Jon's commute in minutes?

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz 1 | Fall 2010)

6. **(10 points)** Given that there was exactly one person arriving at **exactly** 8:20am, what is the probability that this person was Jon?

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz 1 | Fall 2010)

7. **(10 points)** What is the probability that Stephen's commute takes at most as long as Jon's commute?

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz 1 | Fall 2010)

8. **(10 points)** Given that Stephen's commute took at most as long as Jon's, what is the conditional probability that Jon waited 3 minutes for his train?

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz 1 | Fall 2010)

Problem 2. (30 points) For each one of the statements below, give either a proof or a counterexample showing that the statement is not always true.

1. **(10 points)** If events A and B are independent, then the events A and B^c are also independent.
 2. **(10 points)** Let A , B , and C be events associated with a common probabilistic model, and assume that $0 < \mathbf{P}(C) < 1$. Suppose that A and B are conditionally independent given C . Then, A and B are conditionally independent given C^c .

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz 1 | Fall 2010)

(Additional space for Problem 2.2)

3. **(10 points)** Let X and Y be independent random variables. Then, $\text{var}(X + Y) \geq \text{var}(X)$.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Summary of Results for Special Random Variables

Discrete Uniform over $[a, b]$:

$$p_X(k) = \begin{cases} \frac{1}{b-a+1}, & \text{if } k = a, a+1, \dots, b, \\ 0, & \text{otherwise,} \end{cases}$$

$$\mathbf{E}[X] = \frac{a+b}{2}, \quad \text{var}(X) = \frac{(b-a)(b-a+2)}{12}.$$

Bernoulli with Parameter p : (Describes the success or failure in a single trial.)

$$p_X(k) = \begin{cases} p, & \text{if } k = 1, \\ 1-p, & \text{if } k = 0, \end{cases}$$

$$\mathbf{E}[X] = p, \quad \text{var}(X) = p(1-p).$$

Binomial with Parameters p and n : (Describes the number of successes in n independent Bernoulli trials.)

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \dots, n,$$

$$\mathbf{E}[X] = np, \quad \text{var}(X) = np(1-p).$$

Geometric with Parameter p : (Describes the number of trials until the first success, in a sequence of independent Bernoulli trials.)

$$p_X(k) = (1-p)^{k-1} p, \quad k = 1, 2, \dots,$$

$$\mathbf{E}[X] = \frac{1}{p}, \quad \text{var}(X) = \frac{1-p}{p^2}.$$

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz 1 | Fall 2009)

Problem B: (98 points) As a way to practice his probability skills, Bob goes apple picking. The orchard he goes to grows two varieties of apples: gala and honey crisp.

The proportion of the **gala apples** in the orchard is p ($0 < p < 1$), the proportion of the **honey crisp apples** is $1 - p$. The number of apples in the orchard is so large that you can assume that picking a few apples does not change the proportion of the two varieties.

Independent of all other apples, the probability that a randomly picked **gala apple is ripe** is g and the probability that a randomly picked **honey crisp apple is ripe** is h .

1. **(10 points)** Suppose that Bob picks an apple at random (uniformly) and eats it. Find the probability that it was a **ripe gala** apple.

Note: Parts 2 and 3 below can be done independently.

2. Suppose that Bob picks n apples at random (independently and uniformly).
 - (a) **(10 points)** Find the probability that exactly k of those are **gala apples**.
 - (b) Suppose that there are **exactly k gala apples** among the n apples Bob picked. Caleb comes by and gives Bob a **ripe gala** apple to add to his bounty. Bob then picks an apple at random from the $n + 1$ apples and eats it.
 - (i) **(12 points)** What is the probability that it was a **ripe apple**?
 - (ii) **(12 points)** What is the probability that it was a **gala apple if it was ripe**?
 - (c) **(10 points)** Let $n = 20$, and suppose that Bob picked exactly 10 gala apples. What is the probability that the first 10 apples that Bob picked were all gala?
3. Next, Bob tries a different strategy. He starts with a tree of the **gala** variety and picks apples at random from that tree. Once Bob picks an apple off the tree, he carefully examines it to make sure it is ripe. Once he comes across an apple that is not ripe, he moves to **another gala tree**. He does this until he encounters an unripe apple on that second tree. Assume that each tree has a very large, essentially infinite, number of apples.
 - (a) **(10 points)** Let X_i be the **number of apples** Bob picks off the i th tree, ($i = 1, 2$). Write down the PMF, expectation, and variance of X_i .
 - (b) **(12 points)** For $i = 1, 2$, let Y_i be the **total number of ripe apples** Bob picked from the first i trees. Find the expectation and the variance of Y_2 . (Note that $Y_1 = X_1 - 1$ and $Y_2 = (X_1 - 1) + (X_2 - 1)$.)
 - (c) **(12 points)** Find the joint PMF of Y_1 and Y_2 .
 - (d) In the following, answer just “yes” or “no.” (Explanations will not be taken into account in grading.)
 - (i) **(5 points)** Are X_1 and Y_2 independent?
 - (ii) **(5 points)** Are X_2 and Y_1 independent?

Each question is repeated in the following pages. Please write your answer on the appropriate page.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz 1 | Fall 2009)

1. **(10 points)** Suppose that Bob picks an apple at random (uniformly) and eats it. Find the probability that it was a **ripe gala** apple.

Note: Parts 2 and 3 can be done independently.

2. Suppose that Bob picks n apples at random (independently and uniformly).
 - (a) **(10 points)** Find the probability that exactly k of those are **gala apples**.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz 1 | Fall 2009)

- (b) Suppose that there are **exactly k gala apples** among the n apples Bob picked. Caleb comes by and gives Bob a **ripe gala** apple to add to his bounty. Bob then picks an apple at random from the $n + 1$ apples and eats it.
- (i) **(12 points)** What is the probability that it was a **ripe apple**?

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz 1 | Fall 2009)

(ii) **(12 points)** What is the probability that it was a **gala apple if it was ripe?**

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz 1 | Fall 2009)

- (c) **(10 points)** Let $n = 20$, and suppose that Bob picked exactly 10 gala apples. What is the probability that the first 10 apples that Bob picked were all gala?

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz 1 | Fall 2009)

3. Next, Bob tries a different strategy. He starts with a tree of the **gala** variety and picks apples at random from that tree. Once Bob picks an apple off the tree, he carefully examines it to make sure it is ripe. Once he comes across an apple that is not ripe, he moves to **another gala tree**. He does this until he encounters an unripe apple on that second tree. Assume that each tree has a very large, essentially infinite, number of apples.

- (a) **(10 points)** Let X_i be the **number of apples** Bob picks off the i th tree, ($i = 1, 2$). Write down the PMF, expectation, and variance of X_i .

- (b) **(12 points)** For $i = 1, 2$, let Y_i be the **total number of ripe apples** Bob picked from the first i trees. Find the expectation and the variance of Y_2 . (Note that $Y_1 = X_1 - 1$ and $Y_2 = (X_1 - 1) + (X_2 - 1)$.)

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz 1 | Fall 2009)

(c) **(12 points)** Find the joint PMF of Y_1 and Y_2 .

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz 1 | Fall 2009)

(d) In the following, answer just “yes” or “no.” (Explanations will not be taken into account in grading.)

(i) **(5 points)** Are X_1 and Y_2 independent?

(ii) **(5 points)** Are X_2 and Y_1 independent?

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Quiz I Review

Probabilistic Systems Analysis

6.041SC

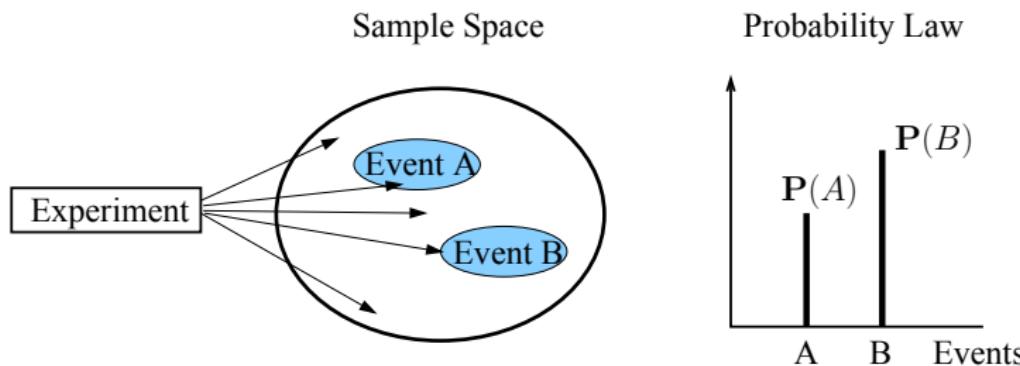
Massachusetts Institute of Technology

Quiz Information

- Content: Chapters 1-2, Lecture 1-7, Recitations 1-7, Psets 1-4, Tutorials 1-3

A Probabilistic Model:

- **Sample Space:** The set of all possible outcomes of an experiment.
- **Probability Law:** An assignment of a nonnegative number $P(E)$ to each event E.



Probability Axioms

Given a sample space Ω :

1. **Nonnegativity:** $P(A) \geq 0$ for each event A
2. **Additivity:** If A and B are disjoint events, then

$$P(A \cup B) = P(A) + P(B)$$

If A_1, A_2, \dots , is a sequence of disjoint events,

$$P(A_1 \cup A_2 \cup \dots) = P(A_1) + P(A_2) + \dots$$

3. **Normalization** $P(\Omega) = 1$

Properties of Probability Laws

Given events A , B and C :

1. If $A \subset B$, then $\mathbf{P}(A) \leq \mathbf{P}(B)$
2. $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B)$
3. $\mathbf{P}(A \cup B) \leq \mathbf{P}(A) + \mathbf{P}(B)$
4. $\mathbf{P}(A \cup B \cup C) = \mathbf{P}(A) + \mathbf{P}(A^c \cap B) + \mathbf{P}(A^c \cap B^c \cap C)$

Discrete Models

- **Discrete Probability Law:** If Ω is finite, then each event $A \subseteq \Omega$ can be expressed as

$$A = \{s_1, s_2, \dots, s_n\} \quad s_i \in \Omega$$

Therefore the probability of the event A is given as

$$\mathbf{P}(A) = \mathbf{P}(s_1) + \mathbf{P}(s_2) + \cdots + \mathbf{P}(s_n)$$

- **Discrete Uniform Probability Law:** If all outcomes are equally likely,

$$\mathbf{P}(A) = \frac{|A|}{|\Omega|}$$

Conditional Probability

- Given an event B with $\mathbf{P}(B) > 0$, the conditional probability of an event $A \subseteq \Omega$ is given as

$$\mathbf{P}(A|B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)}$$

- $\mathbf{P}(A|B)$ is a valid probability law on Ω , satisfying
 - $\mathbf{P}(A|B) \geq 0$
 - $\mathbf{P}(\Omega|B) = 1$
 - $\mathbf{P}(A_1 \cup A_2 \cup \dots | B) = \mathbf{P}(A_1|B) + \mathbf{P}(A_2|B) + \dots$,
where $\{A_i\}_i$ is a set of disjoint events
- $\mathbf{P}(A|B)$ can also be viewed as a probability law on the restricted universe B .

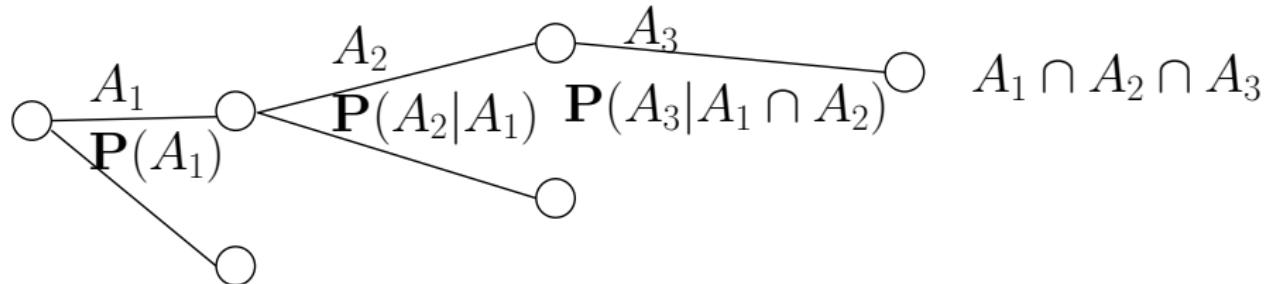
Multiplication Rule

- Let A_1, \dots, A_n be a set of events such that

$$\mathbf{P}\left(\bigcap_{i=1}^{n-1} A_i\right) > 0.$$

Then the joint probability of all events is

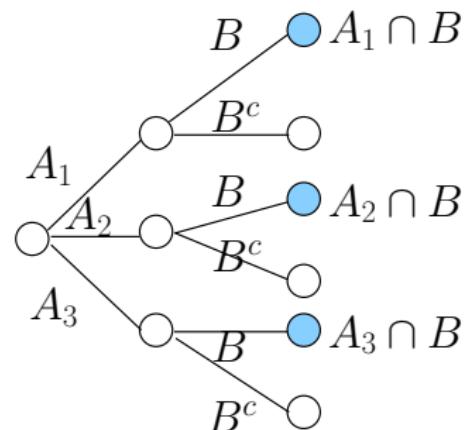
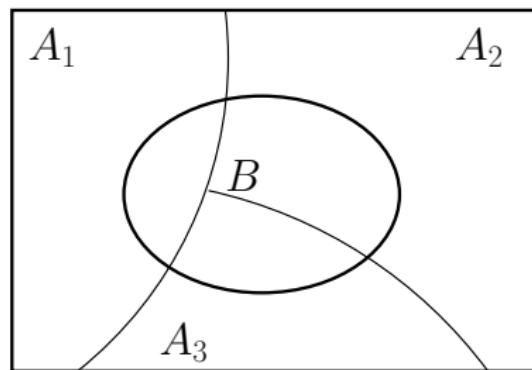
$$\mathbf{P}\left(\bigcap_{i=1}^n A_i\right) = \mathbf{P}(A_1)\mathbf{P}(A_2|A_1)\mathbf{P}(A_3|A_1 \cap A_2) \cdots \mathbf{P}(A_n|\bigcap_{i=1}^{n-1} A_i)$$



Total Probability Theorem

Let A_1, \dots, A_n be disjoint events that partition Ω . If $\mathbf{P}(A_i) > 0$ for each i , then for any event B ,

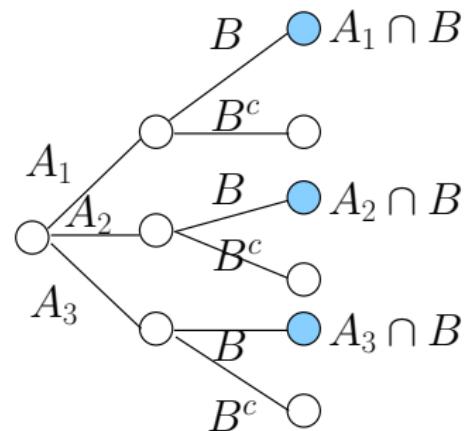
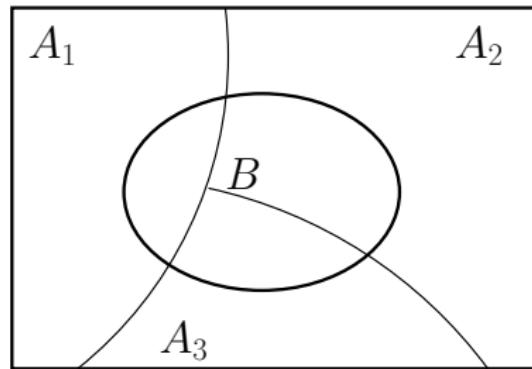
$$\mathbf{P}(B) = \sum_{i=1}^n \mathbf{P}(B \cap A_i) = \sum_{i=1}^n \mathbf{P}(B|A_i)\mathbf{P}(A_i)$$



Bayes Rule

Given a finite partition A_1, \dots, A_n of Ω with $\mathbf{P}(A_i) > 0$,
then for each event B with $\mathbf{P}(B) > 0$

$$\mathbf{P}(A_i|B) = \frac{\mathbf{P}(B|A_i)\mathbf{P}(A_i)}{\mathbf{P}(B)} = \frac{\mathbf{P}(B|A_i)\mathbf{P}(A_i)}{\sum_{i=1}^n \mathbf{P}(B|A_i)\mathbf{P}(A_i)}$$



Independence of Events

- Events A and B are **independent** if and only if

$$\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B)$$

or

$$\mathbf{P}(A|B) = \mathbf{P}(A) \quad \text{if } \mathbf{P}(B) > 0$$

- Events A and B are **conditionally independent** given an event C if and only if

$$\mathbf{P}(A \cap B|C) = \mathbf{P}(A|C)\mathbf{P}(B|C)$$

or

$$\mathbf{P}(A|B \cap C) = \mathbf{P}(A|C) \quad \text{if } \mathbf{P}(B \cap C) > 0$$

- Independence \nLeftrightarrow Conditional Independence.

Independence of a Set of Events

- The events A_1, \dots, A_n are **pairwise independent** if for each $i \neq j$

$$\mathbf{P}(A_i \cap A_j) = \mathbf{P}(A_i)\mathbf{P}(A_j)$$

- The events A_1, \dots, A_n are **independent** if

$$\mathbf{P}\left(\bigcap_{i \in S} A_i\right) = \prod_{i \in S} \mathbf{P}(A_i) \quad \forall S \subseteq \{1, 2, \dots, n\}$$

- Pairwise independence $\not\Rightarrow$ independence, but independence \Rightarrow pairwise independence.

Counting Techniques

- **Basic Counting Principle:** For an m -stage process with n_i choices at stage i ,

$$\# \text{ Choices} = n_1 n_2 \cdots n_m$$

- **Permutations:** k -length sequences drawn from n distinct items without replacement (order is important):

$$\# \text{ Sequences} = n(n - 1) \cdots (n - k + 1) = \frac{n!}{(n-k)!}$$

- **Combinations:** Sets with k elements drawn from n distinct items (order within sets is not important):

$$\# \text{ Sets} = \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Counting Techniques-contd

- **Partitions:** The number of ways to partition an n -element set into r disjoint subsets, with n_k elements in the k^{th} subset:

$$\binom{n}{n_1, n_2, \dots, n_r} = \frac{n!}{n_1! n_2! \cdots n_r!} = \binom{n}{n_1} \binom{n-n_1}{n_2} \cdots \binom{n-n_1-\cdots-n_{r-1}}{n_r}$$

where

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$
$$\sum_{i=1}^r n_i = n$$

Discrete Random Variables

- A **random variable** is a real-valued function defined on the sample space:

$$X : \Omega \rightarrow \mathbb{R}$$

- The notation $\{X = x\}$ denotes an event:

$$\{X = x\} = \{\omega \in \Omega | X(\omega) = x\} \subseteq \Omega$$

- The **probability mass function (PMF)** for the random variable X assigns a probability to each event $\{X = x\}$:

$$p_X(x) = \mathbf{P}(\{X = x\}) = \mathbf{P}(\{\omega \in \Omega | X(\omega) = x\})$$

PMF Properties

- Let X be a random variable and S a countable subset of the real line
- The axioms of probability hold:
 - $p_X(x) \geq 0$
 - $\mathbf{P}(X \in S) = \sum_{x \in S} p_X(x)$
 - $\sum_x p_X(x) = 1$
- If g is a real-valued function, then $Y = g(X)$ is a random variable:

$$\omega \xrightarrow{X} X(\omega) \xrightarrow{g} g(X(\omega)) = Y(\omega)$$

with PMF

$$p_Y(y) = \sum_{x|g(x)=y} P_X(x)$$

Expectation

Given a random variable X with PMF $p_X(x)$:

- $\mathbf{E}[X] = \sum_x x p_X(x)$
- Given a derived random variable $Y = g(X)$:

$$\mathbf{E}[g(X)] = \sum_x g(x) p_X(x) = \sum_y y p_Y(y) = E[Y]$$

$$\mathbf{E}[X^n] = \sum_x x^n p_X(x)$$

- **Linearity** of Expectation: $\mathbf{E}[aX + b] = a\mathbf{E}[X] + b.$

Variance

The expected value of a derived random variable $g(X)$ is

$$\mathbf{E}[g(X)] = \sum_x g(x)p_X(x)$$

The variance of X is calculated as

- $\text{var}(X) = \mathbf{E}[(X - \mathbf{E}[X])^2] = \sum_x (x - \mathbf{E}[X])^2 p_X(x)$
- $\text{var}(X) = \mathbf{E}[X^2] - \mathbf{E}[X]^2$
- $\text{var}(aX + b) = a^2 \text{var}(X)$

Note that $\text{var}(X) \geq 0$

Multiple Random Variables

Let X and Y denote random variables defined on a sample space Ω .

- The **joint PMF** of X and Y is denoted by

$$p_{X,Y}(x,y) = \mathbf{P}(\{X = x\} \cap \{Y = y\})$$

- The **marginal PMFs** of X and Y are given respectively as

$$p_X(x) = \sum_y p_{X,Y}(x,y)$$

$$p_Y(y) = \sum_x p_{X,Y}(x,y)$$

Functions of Multiple Random Variables

Let $Z = g(X, Y)$ be a function of two random variables

- **PMF:**

$$p_Z(z) = \sum_{\substack{(x,y)|g(x,y)=z}} p_{X,Y}(x,y)$$

- **Expectation:**

$$\mathbf{E}[Z] = \sum_{x,y} g(x,y) p_{X,Y}(x,y)$$

- **Linearity:** Suppose $g(X, Y) = aX + bY + c$.

$$\mathbf{E}[g(X, Y)] = a\mathbf{E}[X] + b\mathbf{E}[Y] + c$$

Conditioned Random Variables

- Conditioning X on an event A with $\mathbf{P}(A) > 0$ results in the PMF:

$$p_{X|A}(x) = \mathbf{P}(\{X = x\}|A) = \frac{\mathbf{P}(\{X = x\} \cap A)}{\mathbf{P}(A)}$$

- Conditioning X on the event $Y = y$ with $\mathbf{P}_Y(y) > 0$ results in the PMF:

$$p_{X|Y}(x|y) = \frac{\mathbf{P}(\{X = x\} \cap \{Y = y\})}{\mathbf{P}(\{Y = y\})} = \frac{p_{X,Y}(x,y)}{p_Y(y)}$$

Conditioned RV - contd

- Multiplication Rule:

$$p_{X,Y}(x,y) = p_{X|Y}(x|y)p_Y(y)$$

- Total Probability Theorem:

$$p_X(x) = \sum_{i=1}^n \mathbf{P}(A_i)p_{X|A_i}(x)$$

$$p_X(x) = \sum_y p_{X|Y}(x|y)p_Y(y)$$

Conditional Expectation

Let X and Y be random variables on a sample space Ω .

- Given an event A with $\mathbf{P}(A) > 0$

$$\mathbf{E}[X|A] = \sum_x x p_{X|A}(x)$$

- If $P_Y(y) > 0$, then

$$\mathbf{E}[X|\{Y = y\}] = \sum_x x p_{X|Y}(x|y)$$

- Total Expectation Theorem:** Let A_1, \dots, A_n be a partition of Ω . If $\mathbf{P}(A_i) > 0 \ \forall i$, then

$$\mathbf{E}[X] = \sum_{i=1}^n \mathbf{P}(A_i) \mathbf{E}[X|A_i]$$

Independence

Let X and Y be random variables defined on Ω and let A be an event with $\mathbf{P}(A) > 0$.

- X is independent of A if either of the following hold:

$$p_{X|A}(x) = p_X(x) \quad \forall x$$

$$p_{X,A}(x) = p_X(x)\mathbf{P}(A) \quad \forall x$$

- X and Y are independent if either of the following hold:

$$p_{X|Y}(x|y) = p_X(x) \quad \forall x \forall y$$

$$p_{X,Y}(x,y) = p_X(x)p_Y(y) \quad \forall x \forall y$$

Independence

If X and Y are independent, then the following hold:

- If g and h are real-valued functions, then $g(X)$ and $h(Y)$ are independent.
- $\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y]$ (inverse is not true)
- $\text{var}(X + Y) = \text{var}(X) + \text{var}(Y)$

Given independent random variables X_1, \dots, X_n ,

$$\text{var}(X_1 + X_2 + \dots + X_n) = \text{var}(X_1) + \text{var}(X_2) + \dots + \text{var}(X_n)$$

Some Discrete Distributions

	X	$p_X(k)$	$E[X]$	$var(X)$
Bernoulli	$\begin{cases} 1 & \text{success} \\ 0 & \text{failure} \end{cases}$	$\begin{cases} p & k = 1 \\ 1 - p & k = 0 \end{cases}$	p	$p(1 - p)$
Binomial	Number of successes in n Bernoulli trials	$\binom{n}{k} p^k (1 - p)^{n-k}$ $k = 0, 1, \dots, n$	np	$np(1-p)$
Geometric	Number of trials until first success	$(1 - p)^{k-1} p$ $k = 1, 2, \dots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Uniform	An integer in the interval [a,b]	$\begin{cases} \frac{1}{b-a+1} & k = a, \dots, b \\ 0 & \text{otherwise} \end{cases}$	$\frac{a+b}{2}$	$\frac{(b-a)(b-a+2)}{12}$

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz 1 | Spring 2009)

Question 1

Multiple Choice Questions: **CLEARLY** circle the appropriate choice. Scratch paper is available if needed, though **NO** partial credit will be given for the Multiple Choice.

- a. Which of the following statements is NOT true?
 - (i) If $A \subset B$, then $\mathbf{P}(A) \leq \mathbf{P}(B)$.
 - (ii) If $\mathbf{P}(B) > 0$, then $\mathbf{P}(A|B) \geq \mathbf{P}(A)$.
 - (iii) $\mathbf{P}(A \cap B) \geq \mathbf{P}(A) + \mathbf{P}(B) - 1$.
 - (iv) $\mathbf{P}(A \cap B^c) = \mathbf{P}(A \cup B) - \mathbf{P}(B)$.
- b. We throw n identical balls into m urns at random, where each urn is equally likely and each throw is independent of any other throw. What is the probability that the i -th urn is empty?
 - (i) $\left(1 - \frac{1}{m}\right)^n$
 - (ii) $\left(1 - \frac{1}{n}\right)^m$
 - (iii) $\binom{m}{n} \left(1 - \frac{1}{n}\right)^m$
 - (iv) $\binom{n}{m} \left(\frac{1}{m}\right)^n$
- c. We toss two fair coins simultaneously and independently. If the outcomes of the two coins are the same, we win; otherwise, we lose. Let A be the event that the first coin comes up heads, B be the event that the second coin comes up heads, and C be the event that we win. Which of the following statements is false?
 - (i) Events A and B are independent.
 - (ii) Events A and C are *not* independent.
 - (iii) Events A and B are *not* conditionally independent given C .
 - (iv) The probability of winning is $1/2$.
- d. For a biased coin, the probability of “heads” is $1/3$. Let h be the number of heads in five independent coin tosses. What is the probability $\mathbf{P}(\text{first toss is a head} \mid h = 1 \text{ or } h = 5)$?
 - (i) $\frac{\frac{1}{3}(\frac{2}{3})^4}{\frac{1}{3}(\frac{2}{3})^4 + (\frac{1}{3})^5}$
 - (ii) $\frac{\frac{1}{3}(\frac{2}{3})^4}{\frac{1}{3}(\frac{2}{3})^4 + (\frac{1}{3})^5}$
 - (iii) $\frac{\frac{1}{3}(\frac{2}{3})^4 + (\frac{1}{3})^5}{\frac{1}{3}(\frac{2}{3})^4 + (\frac{1}{3})^5}$
 - (iv) $\frac{1}{5}$

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz 1 | Spring 2009)

e. A well-shuffled deck of 52 cards is dealt evenly to two players (26 cards each). What is the probability that player 1 gets all the aces?

$$(i) \frac{\binom{48}{22} \binom{52}{26}}{\binom{52}{26}}$$

$$(ii) \frac{4 \binom{48}{22} \binom{52}{26}}{\binom{52}{26}}$$

$$(iii) \frac{48! 52!}{22! 26!}$$

$$(iv) \frac{4! \binom{48}{22} \binom{52}{26}}{\binom{52}{26}}$$

f. Suppose X, Y and Z are three independent discrete random variables. Then, X and $Y + Z$ are

- (i) always
- (ii) sometimes
- (iii) never

independent.

g. To obtain a driving licence, Mina needs to pass her driving test. Every time Mina takes a driving test, with probability $1/2$, she will clear the test independent of her past. Mina failed her first test. Given this, let Y be the additional number of tests Mina takes before obtaining a licence. Then,

- (i) $E[Y] = 1$.
- (ii) $E[Y] = 2$.
- (iii) $E[Y] = 0$.

h. Consider two random variables X and Y , each taking values in $\{1, 2, 3\}$. Let their joint PMF be such that for any $1 \leq x, y \leq 3$,

$$P_{X,Y}(x,y) = \begin{cases} 0 & \text{if } (x,y) \in \{(1,3), (2,1), (3,2)\} \\ \text{strictly positive} & \text{otherwise.} \end{cases}$$

Then,

- (i) X and Y can be independent or dependent depending upon the *strictly positive* values.
- (ii) X and Y are always independent.
- (iii) X and Y can never be independent.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz 1 | Spring 2009)

- i. Suppose you play a *matching coins* game with your friend as follows. Both you and your friend have a coin. Each time, you two reveal a side (i.e. H or T) of your coin to each other simultaneously. If the sides match, you WIN a 1 from your friend and if sides do not match then you lose a 1 to your friend. Your friend has a complicated (unknown) strategy in selecting the sides over time. You decide to go with the following simple strategy. Every time, you will toss your unbiased coin independently of everything else, and you will reveal its outcome to your friend (of course, your friend does not know the outcome of your random toss until you reveal it). Then,
- (i) On average, you will lose money to your smart friend.
 - (ii) On average, you will neither lose nor win. That is, your average gain/loss is 0.
 - (iii) On average, you will make money from your friend.
- j. Let $X_i, 1 \leq i \leq 4$ be independent Bernoulli random variable each with mean $p = 0.1$. Let $X = \sum_{i=1}^4 X_i$. That is, X is a Binomial random variable with parameters $n = 4$ and $p = 0.1$. Then,
- (i) $E[X_1|X = 2] = 0.1$.
 - (ii) $E[X_1|X = 2] = 0.5$.
 - (iii) $E[X_1|X = 2] = 0.25$.

Question 2:

Alice and Bob both need to buy a bicycle. The bike store has a stock of four green, three yellow, and two red bikes. Alice randomly picks one of the bikes and buys it. Immediately after, Bob does the same. The sale price of the green, yellow, and red bikes are \$300, \$200 and \$100, respectively.

Let A be the event that Alice bought a green bike, and B be the event that Bob bought a green bike.

a. What is $\mathbf{P}(A)$? What is $\mathbf{P}(A|B)$?

b. Are A and B independent events? Justify your answer.

c. What is the probability that at least one of them bought a green bike?

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz 1 | Spring 2009)

- d. What is the probability that Alice and Bob bought bicycles of different colors?
- e. Given that Bob bought a green bike, what is the expected value of the amount of money spent by Alice?
- f. Let G be the number of green bikes that remain on the store after Alice and Bob's visit. Compute $\mathbf{P}(B|G = 3)$.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz 1 | Spring 2009)

Question 3:

Magic Games Inc. is a store that sells all sorts of fun games. One of its popular products is its magic 4-sided dice. The dice come in pairs; each die can be fair or crooked, and the dice in any pair can function independently or, in some cases, can have magnets inside them that cause them to behave in unpredictable ways when rolled together.

Xavier and Yvonne together buy a pair of dice from this store. Each of them picks a die in the pair; one of them then rolls the two dice together. Let X be the outcome of Xavier's die and Y the outcome of Yvonne's die. The joint PMF of X and Y , $p_{X,Y}(x,y)$, is given by the following figure:

	4	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{10}$
	3	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{10}$	$\frac{1}{20}$
Y	2	$\frac{1}{20}$	$\frac{1}{10}$	$\frac{1}{20}$	$\frac{1}{20}$
	1	$\frac{1}{10}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$
		1	2	3	4
			X		

(a) Find the PMF of the outcome of Xavier's die, $p_X(x)$.

(b) Find the PMF of the outcome of Yvonne's die, $p_Y(y)$.

(c) Are X and Y independent?

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz 1 | Spring 2009)

Zach and Wendy are intrigued by Xavier and Yvonne's dice; they visit the store and buy a pair of dice of their own. Again, each of them picks a die in the pair; one of them then rolls the two dice together. Let Z be the outcome of Zach's die and W the outcome of Wendy's die. The joint PMF of Z and W , $p_{Z,W}(z,w)$, is given by the following figure:

	4	$\frac{1}{12}$	$\frac{1}{24}$	$\frac{1}{12}$	$\frac{1}{24}$
	3	$\frac{1}{12}$	$\frac{1}{24}$	$\frac{1}{12}$	$\frac{1}{24}$
W	2	$\frac{1}{12}$	$\frac{1}{24}$	$\frac{1}{12}$	$\frac{1}{24}$
	1	$\frac{1}{12}$	$\frac{1}{24}$	$\frac{1}{12}$	$\frac{1}{24}$
		1	2	3	4
			Z		

The store also sells a variety of magic coins, some fair and some crooked. Alice buys a coin that on each toss comes up heads with probability $3/4$.

- (d) Wondering whether to buy some dice as well, Alice decides to try out her friends' dice first. She does the following. First, she tosses her coin. If the coin comes up heads, she borrows Xavier and Yvonne's dice pair and rolls the two dice; if the coin comes up tails, she borrows Zach and Wendy's dice pair and rolls those instead. What is the probability that she rolls a double, i.e., that both dice in the pair she rolls show the same number?

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz 1 | Spring 2009)

- (e) Alice has still not made up her mind about the dice. She tries another experiment. First, she tosses her coin. If the coin comes up heads, she takes Xavier and Yvonne's dice pair and rolls the dice repeatedly until she gets a double; if the coin comes up tails, she does the same with Zach and Wendy's dice. What is the expected number of times she will need to roll the dice pair she chooses? (Assume that if a given pair of dice is rolled repeatedly, the outcomes of the different rolls are independent.)
- (f) Alice is bored with the dice and decides to experiment with her coin instead. She tosses the coin until she has seen a total of 11 heads. Let R be the number of tails she sees. Find $\mathbf{E}[R]$. (Assume independent tosses.)

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz 1 | Spring 2009)

- (g) Alice is still playing with her coin. Let A be the event that the second head she sees occurs on the 7th coin toss, and let S be the position of the first head. Find the conditional PMF of S given the event A , $p_{S|A}(s)$.
- (h) Alice's friend Bob buys a coin from the same store that turns out to be fair, i.e., that on any toss comes up heads with probability $1/2$. He tosses the coin repeatedly until he has seen either a total of 11 heads or a total of 11 tails. Let U be the number of times he will need to toss the coin. Find the PMF of U , $p_U(u)$. (Assume independent tosses.)

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz 1 Solutions | Fall 2010)

Quiz 1 Solutions:
October 12, 2010

Problem 1.

1. **(10 points)** Let R_i be the amount of time Stephen spends at the i th red light. R_i is a Bernoulli random variable with $p = 1/3$. The PMF for R_i is:

$$\mathbf{P}_{R_i}(r) = \begin{cases} 2/3, & \text{if } r = 0, \\ 1/3, & \text{if } r = 1, \\ 0, & \text{otherwise.} \end{cases}$$

The expectation and variance for R_i are:

$$\mathbf{E}[R_i] = p = \frac{1}{3},$$

$$\text{var}(R_i) = p(1-p) = \frac{1}{3} \cdot \frac{2}{3} = \frac{2}{9}.$$

Let T_S be the total length of time of Stephen's commute in minutes. Then,

$$T_S = 18 + \sum_{i=1}^5 R_i.$$

T_S is a shifted binomial with $n = 5$ trials and $p = 1/3$. The PMF for T_S is then:

$$\mathbf{P}_{T_S}(k) = \begin{cases} \binom{5}{k-18} \left(\frac{1}{3}\right)^{k-18} \left(\frac{2}{3}\right)^{23-k}, & \text{if } k \in \{18, 19, 20, 21, 22, 23\}, \\ 0, & \text{otherwise.} \end{cases}$$

The expectation and variance for T_S are:

$$\begin{aligned} \mathbf{E}[T_S] &= \mathbf{E}\left[18 + \sum_{i=1}^5 R_i\right] \\ &= \frac{59}{3}. \end{aligned}$$

$$\begin{aligned} \text{var}(T_S) &= \text{var}\left(18 + \sum_{i=1}^5 R_i\right) \\ &= \frac{10}{9}. \end{aligned}$$

2. **(10 points)** Let N be the number of red lights Stephen encountered on his commute. Given that $T_S \leq 19$, then $N = 0$ or $N = 1$. The unconditional probability of $N = 0$ is $\mathbf{P}(N = 0) = (\frac{2}{3})^5$. The unconditional probability of $N = 1$ is $\mathbf{P}(N = 1) = \binom{5}{1}(\frac{2}{3})^4(\frac{1}{3})^1$.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz 1 Solutions | Fall 2010)

To find the conditional expectation, the following conditional PDF is calculated:

$$\mathbf{P}_{N|T_S \leq 19}(n | T_S \leq 19) = \begin{cases} \frac{\left(\frac{2}{3}\right)^5}{\left(\frac{2}{3}\right)^5 + \binom{5}{1}\left(\frac{2}{3}\right)^4\left(\frac{1}{3}\right)^1}, & \text{if } n = 0, \\ \frac{\binom{5}{1}\left(\frac{2}{3}\right)^4\left(\frac{1}{3}\right)^1}{\left(\frac{2}{3}\right)^5 + \binom{5}{1}\left(\frac{2}{3}\right)^4\left(\frac{1}{3}\right)^1}, & \text{if } n = 1, \\ 0, & \text{otherwise,} \end{cases} = \begin{cases} 2/7, & \text{if } n = 0, \\ 5/7, & \text{if } n = 1, \\ 0, & \text{otherwise.} \end{cases}$$

Therefore,

$$\mathbf{E}[N | T_S \leq 19] = \frac{5}{7}.$$

3. **(10 points)** Given that the last red light encountered by Stephen was the fourth light, $R_4 = 1$ and $R_5 = 0$.

We are asked to compute $\text{var}(N | \{R_4 = 1\} \cap \{R_5 = 0\})$. Therefore,

$$\begin{aligned} \text{var}(N | \{R_4 = 1\} \cap \{R_5 = 0\}) &= \text{var}(R_1 + R_2 + R_3 + R_4 + R_5 | \{R_4 = 1\} \cap \{R_5 = 0\}) \\ &= \text{var}(R_1 + R_2 + R_3 + 1 + 0 | \{R_4 = 1\} \cap \{R_5 = 0\}) \\ &= \text{var}(R_1 + R_2 + R_3 + 1) \\ &= \text{var}(R_1 + R_2 + R_3) \\ &= 3\text{var}(R_1) \\ &= \frac{6}{9}. \end{aligned}$$

4. **(10 points)** Under the given condition, the discrete uniform law can be used to compute the probability of interest. There are $\binom{5}{3}$ ways that Stephen can encounter a total of three red lights. There are $\binom{3}{2}$ ways that two out of the first three lights were red. This leaves one additional red light out of the last two lights and there are $\binom{2}{1}$ possible ways that this event can occur. Putting it all together,

$$\mathbf{P}(\text{two of first three lights were red} | \text{total of three red lights}) = \frac{\binom{3}{2}\binom{2}{1}}{\binom{5}{3}} = \frac{3}{5}.$$

5. **(5 points)** Let T_J be the total length of time of Jon's commute in minutes. The PMF of Jon's commute is:

$$\mathbf{P}_{T_J}(\ell) = \begin{cases} \frac{1}{4}, & \text{if } \ell \in \{20, 21, 22, 23\}, \\ 0, & \text{otherwise.} \end{cases}$$

6. **(10 points)** Let A be the event that Jon arrives at work in 20 minutes and let B be the event that exactly one person arrives in 20 minutes.

$$\begin{aligned} \mathbf{P}(A | B) &= \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)} \\ &= \frac{\mathbf{P}(\{T_J = 20\} \cap \{T_S \neq 20\})}{\mathbf{P}(\{T_J = 20\} \cap \{T_S \neq 20\}) + \mathbf{P}(\{T_J \neq 20\} \cap \{T_S = 20\})} \\ &= \frac{\mathbf{P}(T_J = 20)\mathbf{P}(T_S \neq 20)}{\mathbf{P}(T_J = 20)\mathbf{P}(T_S \neq 20) + \mathbf{P}(T_J \neq 20)\mathbf{P}(T_S = 20)}. \end{aligned}$$

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz 1 Solutions | Fall 2010)

Jon arrives at work in 20 minutes (or $T_J = 20$) if he does not have to wait for the train at the station (or $X = 0$). The probability of this event occurring is:

$$\mathbf{P}(T_J = 20) = \mathbf{P}(X = 0) = \frac{1}{4}.$$

Stephen arrives at work in 20 minutes if he encounters 2 red lights. The probability of this event is a binomial probability:

$$\mathbf{P}(T_S = 20) = \binom{5}{2} \left(\frac{1}{3}\right)^2 \left(\frac{2}{3}\right)^3.$$

Thus,

$$\mathbf{P}(A | B) = \frac{\frac{1}{4} \left(1 - \binom{5}{2} \left(\frac{1}{3}\right)^2 \left(\frac{2}{3}\right)^3\right)}{\frac{1}{4} \left(1 - \binom{5}{2} \left(\frac{1}{3}\right)^2 \left(\frac{2}{3}\right)^3\right) + \frac{3}{4} \left(\binom{5}{2} \left(\frac{1}{3}\right)^2 \left(\frac{2}{3}\right)^3\right)}.$$

7. **(10 points)** The probability of interest is $\mathbf{P}(T_S \leq T_J)$. This can be calculated using the total probability theorem by conditioning on the length of Jon's commute or Jon's wait at the station. If Jon's commute is 20 minutes (or $X = 0$), then Stephen can encounter up to 2 red lights to satisfy $T_S \leq T_J$. Similarly if Jon's commute is 21 minutes (or $X = 1$), Stephen can encounter up to 3 red lights and so on.

$$\begin{aligned} \mathbf{P}(T_S \leq T_J) &= \sum_{x=0}^3 \mathbf{P}(T_S \leq T_J | X = x) \mathbf{P}(X = x) \\ &= \frac{1}{4} \sum_{x=0}^3 \sum_{k=0}^{2+x} \binom{5}{k} \left(\frac{1}{3}\right)^k \left(\frac{2}{3}\right)^{5-k} \\ &= 0.9352. \end{aligned}$$

An alternative approach follows. We first compute the joint PMF of the commute times of Stephen and Jon $\mathbf{P}_{T_S, T_J}(k, \ell)$. Because of independence, $\mathbf{P}_{T_S, T_J}(k, \ell) = \mathbf{P}_{T_S}(k) \mathbf{P}_{T_J}(\ell)$.

Therefore,

$$\begin{aligned} \mathbf{P}(T_S \leq T_J) &= \mathbf{P}(T_S = 18) + \mathbf{P}(T_S = 19) + \mathbf{P}(T_S = 20) + \mathbf{P}(\{T_S = 21\} \cap \{T_J \geq 21\}) \\ &\quad + \mathbf{P}(\{T_S = 22\} \cap \{T_J \geq 22\}) + \mathbf{P}(\{T_S = 23\} \cap \{T_J = 23\}) \\ &= \left(\frac{2}{3}\right)^5 + \binom{5}{1} \left(\frac{1}{3}\right)^1 \left(\frac{2}{3}\right)^4 + \binom{5}{2} \left(\frac{1}{3}\right)^2 \left(\frac{2}{3}\right)^3 + \binom{5}{3} \left(\frac{1}{3}\right)^3 \left(\frac{2}{3}\right)^2 \cdot \left(\frac{3}{4}\right) \\ &\quad + \binom{5}{4} \left(\frac{1}{3}\right)^4 \left(\frac{2}{3}\right)^1 \cdot \left(\frac{2}{4}\right) + \left(\frac{1}{3}\right)^5 \cdot \left(\frac{1}{4}\right) \\ &= 0.9352. \end{aligned}$$

8. **(10 points)** We express the conditional probability as such:

$$\mathbf{P}(X = 3 | T_S \leq T_J) = \frac{\mathbf{P}(\{X = 3\} \cap \{T_S \leq T_J\})}{\mathbf{P}(T_S \leq T_J)}.$$

If Jon waited 3 minutes at the train, his commute was 23 minutes and Stephen's commute takes at most as long as Jon's commute since the longest possible commute for Stephen is 23 minutes. Therefore, the numerator in the previous expression is equal to $\mathbf{P}(X = 3) = \frac{1}{4}$. The denominator was computed in the previous part.

$$\begin{aligned}\mathbf{P}(X = 3 \mid T_S \leq T_J) &= \frac{1}{\sum_{x=0}^3 \sum_{k=0}^{2+x} \binom{5}{k} \left(\frac{1}{3}\right)^k \left(\frac{2}{3}\right)^{5-k}} \\ &= 0.2673.\end{aligned}$$

Problem 2.

1. **(10 points) Always True.** We need to show that

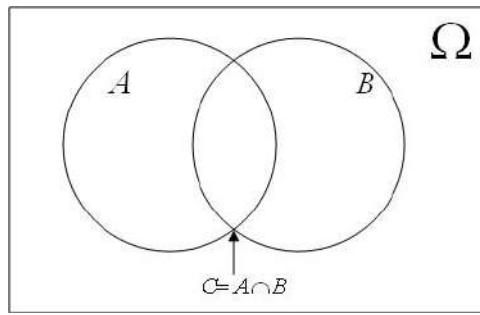
$$\mathbf{P}(A \cap B^c) = \mathbf{P}(A)\mathbf{P}(B^c).$$

We start with expressing $\mathbf{P}(A)$ as $\mathbf{P}(A \cap B) + \mathbf{P}(A \cap B^c)$. Therefore,

$$\begin{aligned}\mathbf{P}(A \cap B^c) &= \mathbf{P}(A) - \mathbf{P}(A \cap B) \\ &= \mathbf{P}(A) - \mathbf{P}(A)\mathbf{P}(B) \\ &= \mathbf{P}(A)(1 - \mathbf{P}(B)) \\ &= \mathbf{P}(A)\mathbf{P}(B^c),\end{aligned}$$

which shows that A and B^c are independent.

2. **(10 points) Not Always True.** Using the diagram below, let $C = A \cap B$ and let $\mathbf{P}(A) > \mathbf{P}(C)$ and let $\mathbf{P}(B) > \mathbf{P}(C)$. The conditional probability $\mathbf{P}(A \cap B \mid C) = 1$. Furthermore, $\mathbf{P}(A \mid C) = 1$ and $\mathbf{P}(B \mid C) = 1$. Since $\mathbf{P}(A \cap B \mid C) = \mathbf{P}(A \mid C)\mathbf{P}(B \mid C)$, A and B are conditionally independent given a third event C . Given C^c , A and B are disjoint which means that A and B are not independent.



The following is an alternative counterexample. Imagine having 3 coins with the following probability of heads: $p = 1/5$, $p = 1/3$ and $p = 2/3$, respectively. Each coin has equal probability of being selected. Let C be the event that you select the coin with $p = 1/5$. Let C^c be the event that you choose one of the other two coins. Let A be the event that the first coin toss results in heads. Let B be the event that the second coin toss results in heads. For a given coin, the tosses are independent such that:

$$\mathbf{P}(B \mid A \cap C) = \mathbf{P}(B \mid C).$$

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz 1 Solutions | Fall 2010)

Given C^c , A and B are not independent since we can have either the $p = 1/3$ coin or the $p = 2/3$ coin. Knowing A changes our beliefs of the result of the second coin toss.

$$\begin{aligned}\mathbf{P}(B \mid A \cap C^c) &= \frac{B \cap A \cap C^c}{A \cap C^c} \\ &= \frac{\frac{1}{3} \left(\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2 \right)}{\frac{1}{3} \left(\frac{1}{3} + \frac{2}{3} \right)} \\ &= \frac{5}{9}.\end{aligned}$$

However,

$$\begin{aligned}\mathbf{P}(B \mid C^c) &= \frac{\mathbf{P}(B \cap C^c)}{\mathbf{P}(C^c)} \\ &= \frac{\frac{1}{3} \left(\frac{1}{3} + \frac{2}{3} \right)}{\frac{2}{3}} \\ &= \frac{1}{2}.\end{aligned}$$

As shown, $\mathbf{P}(B \mid A \cap C^c) \neq \mathbf{P}(B \mid C^c)$.

3. **(10 points) Always True.** Using independence of X and Y , $\text{var}(X + Y) = \text{var}(X) + \text{var}(Y)$. Since variance is always non-negative, $\text{var}(X) + \text{var}(Y) \geq \text{var}(X)$.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Recitation 10
October 12, 2010

- **Question 1.** The two parts of this question are about identities for a probabilistic model with sample space Ω , events A and B , and discrete random variable X . Any time conditioning on an event is indicated, the event has positive probability. An identity is *true* when it holds without any additional restrictions; it is *false* when there is any counterexample.

1.1. Which **one** of the following statements is **true**?

- (a) $\mathbf{P}(A \cap B)$ may be larger than $\mathbf{P}(A)$.
- (b) The variance of X may be larger than the variance of $2X$.
- (c) If $A^c \cap B^c = \emptyset$, then $\mathbf{P}(A \cup B) = 1$.
- (d) If $A^c \cap B^c = \emptyset$, then $\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B)$.
- (e) If $\mathbf{P}(A) > 1/2$ and $\mathbf{P}(B) > 1/2$, then $\mathbf{P}(A \cup B) = 1$.

1.2. Which **one** of the following statements is **true**?

- (a) If $\mathbf{E}[X] = 0$, then $\mathbf{P}(X > 0) = \mathbf{P}(X < 0)$.
- (b) $\mathbf{P}(A) = \mathbf{P}(A | B) + \mathbf{P}(A | B^c)$
- (c) $\mathbf{P}(B | A) + \mathbf{P}(B | A^c) = 1$
- (d) $\mathbf{P}(B | A) + \mathbf{P}(B^c | A^c) = 1$
- (e) $\mathbf{P}(B | A) + \mathbf{P}(B^c | A) = 1$

- **Question 2.**

Provide **clear reasoning**; partial credit is possible

Heather and Taylor play a game using independent tosses of an unfair coin. A head comes up on any toss with probability p , where $0 < p < 1$. The coin is tossed repeatedly until either the second time head comes up, in which case Heather wins; or the second time tail comes up, in which case Taylor wins. Note that a full game involves 2 or 3 tosses.

- 2.1. Consider a probabilistic model for the game in which the outcomes are the sequences of heads and tails in a full game. Provide a list of the outcomes and their probabilities of occurring.
- 2.2. What is the probability that Heather wins the game?
- 2.3. What is the conditional probability that Heather wins the game given that head comes up on the first toss?
- 2.4. What is the conditional probability that head comes up on the first toss given that Heather wins the game?

• **Question 3.**

Provide **clear reasoning**; partial credit is possible

A casino game using a **fair** 4-sided die (with labels 1, 2, 3, and 4) is offered in which a **basic game** has 1 or 2 die rolls:

- If the first roll is a 1, 2, or 3, the player wins the amount of the die roll, in dollars, and the game is over.
- If the first roll is a 4, the player wins \$2 and the amount of a second (“bonus”) die roll in dollars.

Let X be the payoff in dollars of the basic game.

- 3.1. Find the PMF of X , $p_X(x)$.
- 3.2. Find $\mathbf{E}[X]$.
- 3.3. Find the conditional PMF of the result of the first die roll given that $X = 3$. (Use a reasonable notation that you define explicitly.)
- 3.4. Now consider an **extended game** that can have any number of bonus rolls. Specifically:
 - * Any roll of a 1, 2, or 3 results in the player winning the amount of the die roll, in dollars, and the termination of the game.
 - * Any roll of a 4 results in the player winning \$2 and continuation of the game.

Let Y denote the payoff in dollars of the extended game. Find $\mathbf{E}[Y]$.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Recitation 10 Solutions
(6.041/6.431 Spring 2010 Quiz 1 Solutions)

Question 1

1.1. Which **one** of the following statements is **true**?

- (a) $\mathbf{P}(A \cap B)$ may be larger than $\mathbf{P}(A)$.
- (b) The variance of X may be larger than the variance of $2X$.
- (c) If $A^c \cap B^c = \emptyset$, then $\mathbf{P}(A \cup B) = 1$.
- (d) If $A^c \cap B^c = \emptyset$, then $\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B)$.
- (e) If $\mathbf{P}(A) > 1/2$ and $\mathbf{P}(B) > 1/2$, then $\mathbf{P}(A \cup B) = 1$.

Answer: (c) is true because $A \cup B = (A^c \cap B^c)^c = \emptyset^c = \Omega$.

1.2. Which **one** of the following statements is **true**?

- (a) If $\mathbf{E}[X] = 0$, then $\mathbf{P}(X > 0) = \mathbf{P}(X < 0)$.
- (b) $\mathbf{P}(A) = \mathbf{P}(A | B) + \mathbf{P}(A | B^c)$
- (c) $\mathbf{P}(B | A) + \mathbf{P}(B | A^c) = 1$
- (d) $\mathbf{P}(B | A) + \mathbf{P}(B^c | A^c) = 1$
- (e) $\mathbf{P}(B | A) + \mathbf{P}(B^c | A) = 1$

Answer: (e) is true because B and B^c partition Ω .

Question 2

Heather and Taylor play a game using independent tosses of an unfair coin. A head comes up on any toss with probability p , where $0 < p < 1$. The coin is tossed repeatedly until either the second time head comes up, in which case Heather wins; or the second time tail comes up, in which case Taylor wins. Note that a full game involves 2 or 3 tosses.

2.1. Consider a probabilistic model for the game in which the outcomes are the sequences of heads and tails in a full game. Provide a list of the outcomes and their probabilities of occurring.

Because of the independence of the coin tosses, the outcomes and their probabilities are as follows:

HH	p^2
HTH	$p^2(1-p)$
HTT	$p(1-p)^2$
THH	$p^2(1-p)$
THT	$p(1-p)^2$
TT	$(1-p)^2$

2.2. What is the probability that Heather wins the game?

The event of Heather winning is {HH, HTH, THH}. Adding the probabilities of the outcomes in this event gives $p^2 + p^2(1-p) + p^2(1-p) = p^2(3 - 2p)$.

- 2.3. What is the conditional probability that Heather wins the game given that head comes up on the first toss?

$$\begin{aligned}
 P(\{\text{Heather wins}\} \mid \{\text{first toss H}\}) &= \frac{P(\{\text{Heather wins}\} \cap \{\text{first toss H}\})}{P(\{\text{first toss H}\})} \\
 &= \frac{P(\{\text{HH, HTH}\})}{P(\{\text{first toss H}\})} \\
 &= \frac{p^2 + p^2(1-p)}{p} = p(2-p)
 \end{aligned}$$

- 2.4. What is the conditional probability that head comes up on the first toss given that Heather wins the game?

$$\begin{aligned}
 P(\{\text{first toss H}\} \mid \{\text{Heather wins}\}) &= \frac{P(\{\text{first toss H}\} \cap \{\text{Heather wins}\})}{P(\{\text{Heather wins}\})} \\
 &= \frac{P(\{\text{HH, HTH}\})}{P(\{\text{Heather wins}\})} \\
 &= \frac{p^2 + p^2(1-p)}{p^2(3-2p)} = \frac{2-p}{3-2p}
 \end{aligned}$$

Question 3

A casino game using a **fair** 4-sided die (with labels 1, 2, 3, and 4) is offered in which a **basic game** has 1 or 2 die rolls:

- If the first roll is a 1, 2, or 3, the player wins the amount of the die roll, in dollars, and the game is over.
- If the first roll is a 4, the player wins \$2 and the amount of a second (“bonus”) die roll in dollars.

Let X be the payoff in dollars of the basic game.

- 3.1. Find the PMF of X , $p_X(x)$.

Define a probabilistic model in which the outcomes are the sequences of rolls in a full game. The outcomes, their probabilities, and the resulting values of X are as follows:

ω	$P(\{\omega\})$	$X(\omega)$
(1)	1/4	1
(2)	1/4	2
(3)	1/4	3
(4, 1)	1/16	3
(4, 2)	1/16	4
(4, 3)	1/16	5
(4, 4)	1/16	6

By gathering the probabilities of the possible values for X , we obtain

$$p_X(x) = \begin{cases} 1/4, & \text{for } x = 1, 2; \\ 5/16, & \text{for } x = 3; \\ 1/16, & \text{for } x = 4, 5, 6; \\ 0, & \text{otherwise.} \end{cases}$$

3.2. Find $\mathbf{E}[X]$.

It does not take too much arithmetic to compute $\mathbf{E}[X]$ using the PMF computed in the previous part. A more elegant solution is to use the total expectation theorem. Let A be the event that the first roll is a 4. Then

$$\mathbf{E}[X] = \underbrace{\mathbf{P}(A)}_{1/4} \underbrace{\mathbf{E}[X | A]}_{4.5} + \underbrace{\mathbf{P}(A^c)}_{3/4} \underbrace{\mathbf{E}[X | A^c]}_2 = \frac{21}{8},$$

where $\mathbf{E}[X | A] = 4.5$ because the conditional distribution is uniform on $\{3, 4, 5, 6\}$; and $\mathbf{E}[X | A^c] = 2$ because the conditional distribution is uniform on $\{1, 2, 3\}$.

3.3. Find the conditional PMF of the result of the first die roll given that $X = 3$. (Use a reasonable notation that you define explicitly.)

Let Z be the result of the first die roll, and let $B = \{X = 3\}$. By definition of conditioning,

$$p_{Z|B}(z) = \frac{\mathbf{P}(\{Z = z\} \cap B)}{\mathbf{P}(B)}.$$

By using values tabulated above,

$$p_{Z|B}(z) = \begin{cases} 4/5, & \text{for } z = 3; \\ 1/5, & \text{for } z = 4; \\ 0, & \text{otherwise.} \end{cases}$$

3.4. Now consider an **extended game** that can have any number of bonus rolls. Specifically:

- Any roll of a 1, 2, or 3 results in the player winning the amount of the die roll, in dollars, and the termination of the game.
- Any roll of a 4 results in the player winning \$2 and continuation of the game.

Let Y denote the payoff in dollars of the extended game. Find $\mathbf{E}[Y]$.

One could explicitly find the PMF of Y , but this is unnecessarily messy. Instead, let L be the payoff of the last roll and let W be the payoff of all of the earlier rolls. Then $Y = W + L$ by construction, and $\mathbf{E}[Y] = \mathbf{E}[W] + \mathbf{E}[L]$.

The last roll is uniformly distributed on $\{1, 2, 3\}$, so $\mathbf{E}[L] = 2$. The winnings on earlier rolls is $2(N - 1)$ where N is the number of rolls in the game. Since termination of the game can be seen as “success” on a Bernoulli trial with success probability of $3/4$, N has the geometric distribution with parameter $3/4$. Thus,

$$\mathbf{E}[W] = \mathbf{E}[2(N - 1)] = 2\mathbf{E}[N] - 2 = 2 \cdot \frac{4}{3} - 2 = \frac{2}{3}.$$

Combining the calculations,

$$\mathbf{E}[Y] = \mathbf{E}[W] + \mathbf{E}[L] = \frac{2}{3} + 2 = \frac{8}{3}.$$

(Many other methods of solution are possible.)

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013 Transcript – Recitation: A Mixed Distribution Example

In this video, we'll look at an example in which we compute the expectation and cumulative density function of a mixed random variable. The problem is as follows. Al arrives at some bus stand or taxi stand at a given time-- let's say time t equals 0. He finds a taxi waiting for him with probability $2/3$ in which he takes it. Otherwise, he takes the next arriving taxi or bus. The time that the next taxi arrives between 0 and 10 minutes, and it's uniformly distributed. The next bus leaves exactly in 5 minutes. So the question is, if X is Al's waiting time, what is the CDF and expectation of X ?

So one way to view this problem that's convenient is the tree structure. So I've drawn it for you here in which the events of interest are B_1 , B_2 , and B_3 , B_1 being Al catches the waiting taxi, B_2 being Al catches the next taxi, which arrives between 0 and 5 minutes, and B_3 being Al catches the bus at the time t plus 5. Notice that these three events are disjoint. So Al catching the waiting taxi means he can't catch the bus or the next arriving taxi. And it also covers the entire set of outcomes. So, in fact, B_1 , B_2 , and B_3 are a partition.

So let's look at the relevant probabilities. Whether or not B_1 happens depends on whether or not the taxi's waiting for Al. So if the taxi is waiting for him, which happens with $2/3$ probability, B_1 happens. Otherwise, with $1/3$ probability, we see whether or not a taxi is going to arrive between 0 and 5 minutes. If it arrives, which is going to happen with what probability? Well, we know that the next taxi is going to arrive between 0 and 10 minutes uniform. It's a uniform distribution. And so half the mass is going to be between 0 and 5. And the other half is going to be between 5 and 10. And so this is going to be $1/2$ and $1/2$.

And let's look at what X looks like. If B_1 happens, Al isn't waiting at all, so x is going to be equal to 0. If B_3 happens, which is the other easy case, Al's going to be waiting for 5 minutes exactly. And if B_2 happens, well, it's going to be some value between 0 and 5.

We can actually draw the density, so let's see if we can do that here. The original next taxi was uniformly distributed between 0 and 10. But now, we're told two pieces of information. We're told that B_2 happens, which means that there's no taxi waiting, and the next taxi arrives between 0 and 5 minutes. Well, the fact that there was no taxi waiting has no bearing on that density. But the fact that the next taxi arrives between 0 and 5 does make a difference, because the density then is going to be definitely 0 in any region outside 0 and 5.

Now, the question is, how is it going to look between 0 and 5? Well, it's not going to look crazy. It's not going to look like something different. It's simply going to be a scale version of the original density between 0 and 5. You can verify this by looking at the actual formula for when you condition events on a random variable. Here, it's going to be $1/5$ in order for this to integrate out to 1. And now we can jump right into figuring out the expectation.

Now, notice that X is actually a mixed random variable? What does that mean? Well, X either takes on values according to either a discrete probability law or a continuous one. So if B_1

happens, for example, X is going to be exactly equal to 0 with probability 1, which is a discrete probability problem. On the other hand, if B_2 happens, then the value of X depends on the density, which is going to be continuous. So X is going to be a continuous random variable here.

So how do you define an expectation in this case? Well, you can do it so that it satisfies the total expectation theorem, which means that the expectation of X is the probability of B_1 times the expectation given B_1 plus the probability of B_2 times the expectation given B_2 plus the probability of B_3 times the expectation given B_3 . So this will satisfy the total expectation theorem.

So the probability of B_1 is going to be exactly $2/3$. It's simply the probability of a taxi waiting for A_1 . The expected value of X -- well, when B_1 happens, X is going to be exactly equal to 0. So the expected value is also going to be 0. The probability of B_2 happening is the probability of a taxi not being there times the probability of a taxi arriving between 0 and 5. It's going to be $1/3$ times $1/2$.

And the expected value of X given B_2 is going to be the expected value of this density. The expected value of this density is the midpoint between 0 and 5. And so it's going to be $5/2$. And the probability of B_3 is going to be $1/3$ times $1/2$. Finally, the expected value of X given B_3 . Well, when B_3 happens, X is going to be exactly equal to 5. So the expected value is also going to be 5. Now we're left with $5/12$ plus $5/6$, which is going to be $15/12$. And we can actually fill that in here so that we can clear up the board to do the other part.

Now we want to compute the CDF of X . Well, what is the CDF? Well, the CDF of X is going to be equal to the probability that the random variable X is less than or equal to some little x . It's a constant [INAUDIBLE].

Before we jump right in, let's try to understand what's the form of the CDF. And let's consider some interesting cases. You know that the random variable X , the waiting time, is going to be somewhere between 0 and 5, right? So let's consider what happens if little x is going to be less than 0. That's basically saying, what's the probability of the random variable X being less than some number that's less than 0? Waiting time can't be negative, so the probability of this is going to be 0.

Now, what if X is between equaling 0 and strictly less than 5? In that case, either X can fall between 0 and 5 according to this case, in the case of B_2 , or X can be exactly equal to 0. It's not clear. So let's do that later. Let's fill that in later.

What about if x is greater than or equal to 5? Little x , right? That's the probability that the random variable X is less than some number that's bigger than or equal to 5. The waiting time X , the random variable, is definitely going to be less than or equal to 5, so the probability of this is going to be 1.

So now this case. How do we do it? Well, let's try to use a similar kind of approach that we did for the expected value and use the total probability theorem in this case. So let's try to review

this. First of all, let's assume that this is true, that little x is between 0 and 5, including 0. And let's use the total probability theorem, and use the partitions B_1 , B_2 , and B_3 .

So what's the probability of B_1 ? It's the probability that A_1 catches waiting taxi, which happens with probability $2/3$. What's the probability that the random variable X , which is less than or equal to little x under this condition, when B_1 happens? Well, if B_1 happens, then random variable X is going to be exactly equal to 0, right? So in that case, it's definitely going to be less than or equal to any value of x , including 0. So the probability will be 1.

What's the probability that B_2 happens now? The probability that B_2 happens is $1/3$ times $1/2$, as we did before. And the probability that the random variable X is less than or equal to little x when B_2 happens. Well, if B_2 happens, this is your density. And this is our condition. And so x is going to be somewhere in between these spots.

And we'd like to compute what's the probably that random variable X is less than or equal to little x . So we want this area. And that area is going to have height of $1/5$ and width of x . And so the area's going to be $1/5$ times x .

And finally, the probability that B_3 happens is going to be $1/3$ times $1/2$ again times the probability that the random variable X is less than or equal to little x given B_3 . Well, when B_3 happens, X is going to be exactly 5 as a random variable. But little x , you know-- we're assuming in this condition-- is going to be between 0 and 5, but strictly less than 5. So there's no way that if the random variable X is 5 and this is strictly less than 5, this is going to be true. And so that probability will be 0.

So we're now left with $2/3$ plus $1/30$. And now we can fill this in. $2/3$ plus $1/30 x$. And this is our CDF.

So now we've finished the problem, computed the expected value here and then the CDF here, and this was a great illustration of how you would do so for a mixed random variable.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013

Transcript – Recitation: Calculating a Cumulative Distribution Function (CDF)

Hi. In this problem, we'll get some practice working with PDFs and also using PDFs to calculate CDFs. So the PDF that we're given in this problem is here. So we have a random variable, z , which is a continuous random variable. And we're told that the PDF of this random variable, z , is given by γz times $1 + z^2$ in the range of z between negative 2 and 1. And outside of this range, it's 0.

All right, so first thing we need to do and the first part of this problem is we need to figure out what γ is because it's not really a fully specified PDF yet. We need to figure out exactly what the value γ is. And how do we do that? Well, we've done analogous things before for the discrete case.

So the tool that we use is that the PDF must integrate to 1. So in the discrete case, the analogy was that the PMF had to sum to 1. So what do we know? We know that when you integrate this PDF from negative infinity to infinity, $f_z(z)$, it has to equal 1.

All right, so what do we do now? Well, we know what the PDF is-- partially, except for γ -- so let's plug that in. And the first thing that we'll do is we'll simplify this because we know that the PDF is actually only non-zero in the range negative 2 to 1. So instead of integrating from negative infinity to infinity, we'll just integrate from negative 2 to 1.

And now let's plug in this γz times $1 + z^2$ dz . And now the rest of the problem is just applying calculus and integrating this. So let's just go through that process.

So we get $z + \frac{1}{3}z^3$ from minus 2 to 1. And now we'll plug in the limits. And we get γ , and that's $1 + \frac{1}{3} - (-2 + \frac{1}{3})$.

And then if we add this all up, you get $\frac{4}{3} + 2 + \frac{8}{3}$, which will give you 6. So what we end up with in the end is that 1 is equal to 6 γ . So what does that tell us? That tells us that, in this case, γ is $1/6$.

OK, so we've actually figured out what this PDF really is. And let's just substitute that in. So we know what γ is. So it's $1/6$.

So from this PDF, we can calculate anything that we want to. This PDF, basically, fully specifies everything that we need to know about this random variable, z . And one of the things that we can calculate from the PDF is the CDF. So the next part of the problem asks us to calculate the CDF.

So remember the CDF, we use capital F . And the definition is that you integrate from negative infinity to this z . And what do you integrate? You integrate the PDF. And all use some dummy variable, y , here in the integration.

So what is it really doing? It's basically just taking the PDF and taking everything to the left of it. So another way to think about this-- this is the probability that the random variable is less than or equal to some little z . It's just accumulating probability as you go from left to right.

So the hardest part about calculating the CDFs, really, is actually just keeping track of the ranges, because unless the PDF is really simple, you'll have cases where the PDF could be 0 in some ranges and non-zero in other ranges. And then what you really have to keep track of is where those ranges are and where you actually have non-zero probability.

So in this case, we actually break things down into three different ranges because this PDF actually looks something like this. So it's non-zero between negative 2 and 1, and it's 0 everywhere else. So then what that means is that our job is a little simpler because everything to the left of negative 2, the CDF will be 0 because there's no probability density to the left.

And then everything to the right of 1, well we've accumulated all the probability in the PDF because we know that when you integrate from negative 2 to 1, you capture everything. So anything to the right of 1, the CDF will be 1. So the only hard part is calculating what the CDF is in this intermediate range, between negative 2 and 1.

So let's do that case first-- so the case of z is between negative 2 and 1. So what is the CDF in that case? Well, the definition is to integrate from negative infinity to z .

But we know that everything to the left of negative 2, there's no probability density. So we don't need to include that. So we can actually change this lower limit to negative 2. And the upper limit is wherever this z is.

So that becomes our integral. And the inside is still the PDF. So let's just plug that in. We know that it's $1/6$ plus-- we'll make this y squared-- by. And now it's just calculus again. And in fact, it's more or less the same integral, so what we get is y plus $1/3$ y cubed from negative 2 to z .

Notice the only thing that's different here is that we're integrating from negative 2 to z instead of negative 2 to 1. And when we calculate this out, what we get is z plus $1/3$ z cubed minus 2 plus $1/3$ minus 2 cubed, which gives us $1/6$ z plus $1/3$ z cubed plus 2 plus $8/3$ gives us $14/3$. So that actually is our CDF between the range of negative 2 to 1.

So for full completeness, let's actually write out the entire CDF, because there's two other parts in the CDF. So the first part is that it's 0 if z is less than negative 2. And it's 1 if z is greater than 1.

And in between, it's this thing that we've just calculated. So it's $1/6$ z plus $1/3$ z cubed plus $14/3$ if z is between minus 2 and 1. So that is our final answer.

So the main point of this problem was to drill a little bit more the concepts of PDFs and CDFs. So for the PDF, the important thing to remember is that in order to be a valid PDF, the PDF has to integrate to 1. And you can use that fact to help you calculate any unknown constants in the PDF.

And then to calculate the CDF, it's just integrating the PDF from negative infinity to whatever point that you want to cut off at. And the tricky part, as I said earlier, was really just keeping track of the ranges. In this case, we've broke it down into three ranges. If we had a slightly more complicated PDF, then you would have to keep track of even more ranged. All right, so I hope that was helpful, and we'll see you next time.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013 Transcript – Lecture 8

The following content is provided under a Creative Commons license. Your support will help MIT OpenCourseWare continue to offer high-quality educational resources for free. To make a donation or view additional materials from hundreds of MIT courses, visit MIT OpenCourseWare at ocw.mit.edu.

JOHN TSITSIKLIS: OK. We can start. Good morning.

So we're going to start now a new unit. For the next couple of lectures, we will be talking about continuous random variables. So this is new material which is not going to be in the quiz. You are going to have a long break next week without any lecture, just a quiz and recitation and tutorial.

So what's going to happen in this new unit? Basically, we want to do everything that we did for discrete random variables, reintroduce the same sort of concepts but see how they apply and how they need to be modified in order to talk about random variables that take continuous values. At some level, it's all the same. At some level, it's quite a bit harder because when things are continuous, calculus comes in. So the calculations that you have to do on the side sometimes need a little bit more thinking.

In terms of new concepts, there's not going to be a whole lot today, some analogs of things we have done. We're going to introduce the concept of cumulative distribution functions, which allows us to deal with discrete and continuous random variables, all of them in one shot. And finally, introduce a famous kind of continuous random variable, the normal random variable.

OK, so what's the story? Continuous random variables are random variables that take values over the continuum. So the numerical value of the random variable can be any real number. They don't take values just in a discrete set.

So we have our sample space. The experiment happens. We get some omega, a sample point in the sample space. And once that point is determined, it determines the numerical value of the random variable.

Remember, random variables are functions on the sample space. You pick a sample point. This determines the numerical value of the random variable. So that numerical value is going to be some real number on that line.

Now we want to say something about the distribution of the random variable. We want to say which values are more likely than others to occur in a certain sense. For example, you may be interested in a particular event, the event that the random variable takes values in the interval from a to b. And we want to say something about the probability of that event.

In principle, how is this done? You go back to the sample space, and you find all those outcomes for which the value of the random variable happens to be in that interval. The probability that the

random variable falls here is the same as the probability of all outcomes that make the random variable to fall in there. So in principle, you can work on the original sample space, find the probability of this event, and you would be done.

But similar to what happened in chapter 2, we want to kind of push the sample space in the background and just work directly on the real axis and talk about probabilities up here. So we want now a way to specify probabilities, how they are bunched together, or arranged, along the real line. So what did we do for discrete random variables? We introduced PMFs, probability mass functions. And the way that we described the random variable was by saying this point has so much mass on top of it, that point has so much mass on top of it, and so on.

And so we assigned a total amount of 1 unit of probability. We assigned it to different masses, which we put at different points on the real axis. So that's what you do if somebody gives you a pound of discrete stuff, a pound of mass in little chunks. And you place those chunks at a few points.

Now, in the continuous case, this total unit of probability mass does not sit just on discrete points but is spread all over the real axis. So now we're going to have a unit of mass that spreads on top of the real axis. How do we describe masses that are continuously spread?

The way we describe them is by specifying densities. That is, how thick is the mass that's sitting here? How dense is the mass that's sitting there? So that's exactly what we're going to do. We're going to introduce the concept of a probability density function that tells us how probabilities accumulate at different parts of the real axis.

So here's an example or a picture of a possible probability density function. What does that density function kind of convey intuitively? Well, that these x 's are relatively less likely to occur. Those x 's are somewhat more likely to occur because the density is higher.

Now, for a more formal definition, we're going to say that a random variable X is said to be continuous if it can be described by a density function in the following sense. We have a density function. And we calculate probabilities of falling inside an interval by finding the area under the curve that sits on top of that interval.

So that's sort of the defining relation for continuous random variables. It's an implicit definition. And it tells us a random variable is continuous if we can calculate probabilities this way. So the probability of falling in this interval is the area under this curve. Mathematically, it's the integral of the density over this particular interval. If the density happens to be constant over that interval, the area under the curve would be the length of the interval times the height of the density, which sort of makes sense.

Now, because the density is not constant but it kind of moves around, what you need is to write down an integral. Now, this formula is very much analogous to what you would do for discrete random variables. For a discrete random variable, how do you calculate this probability?

You look at all x 's in this interval. And you add the probability mass function over that range. So just for comparison, this would be the formula for the discrete case-- the sum over all x 's in the interval from a to b over the probability mass function. And there is a syntactic analogy that's happening here and which will be a persistent theme when we deal with continuous random variables.

Sums get replaced by integrals. In the discrete case, you add. In the continuous case, you integrate. Mass functions get replaced by density functions. So you can take pretty much any formula from the discrete case and translate it to a continuous analog of that formula, as we're going to see.

OK. So let's take this now as our model. What is the probability that the random variable takes a specific value if we have a continuous random variable? Well, this would be the case. It's a case of a trivial interval, where the two end points coincide. So it would be the integral from a to itself. So you're integrating just over a single point.

Now, when you integrate over a single point, the integral is just 0. The area under the curve, if you're only looking at a single point, it's 0. So big property of continuous random variables is that any individual point has 0 probability.

In particular, when you look at the value of the density, the density does not tell you the probability of that point. The point itself has 0 probability. So the density tells you something a little different. We are going to see shortly what that is.

Before we get there, can the density be an arbitrary function? Almost, but not quite. There are two things that we want. First, since densities are used to calculate probabilities, and since probabilities must be non-negative, the density should also be non-negative. Otherwise you would be getting negative probabilities, which is not a good thing. So that's a basic property that any density function should obey.

The second property that we need is that the overall probability of the entire real line should be equal to 1. So if you ask me, what is the probability that x falls between minus infinity and plus infinity, well, we are sure that x is going to fall in that range. So the probability of that event should be 1. So the probability of being between minus infinity and plus infinity should be 1, which means that the integral from minus infinity to plus infinity should be 1. So that just tells us that there's 1 unit of total probability that's being spread over our space.

Now, what's the best way to think intuitively about what the density function does? The interpretation that I find most natural and easy to convey the meaning of a density is to look at probabilities of small intervals. So let us take an x somewhere here and then x plus delta just next to it. So delta is a small number.

And let's look at the probability of the event that we get a value in that range. For continuous random variables, the way we find the probability of falling in that range is by integrating the density over that range. So we're drawing this picture. And we want to take the area under this curve.

Now, what happens if delta is a fairly small number? If delta is pretty small, our density is not going to change much over that range. So you can pretend that the density is approximately constant. And so to find the area under the curve, you just take the base times the height.

And it doesn't matter where exactly you take the height in that interval, because the density doesn't change very much over that interval. And so the integral becomes just base times the height. So for small intervals, the probability of a small interval is approximately the density times delta. So densities essentially give us probabilities of small intervals.

And if you want to think about it a little differently, you can take that delta from here and send it to the denominator there. And what this tells you is that the density is probability per unit length for intervals of small length. So the units of density are probability per unit length.

Densities are not probabilities. They are rates at which probabilities accumulate, probabilities per unit length. And since densities are not probabilities, they don't have to be less than 1.

Ordinary probabilities always must be less than 1. But density is a different kind of thing. It can get pretty big in some places. It can even sort of blow up in some places. As long as the total area under the curve is 1, other than that, the curve can do anything that it wants.

Now, the density prescribes for us the probability of intervals. Sometimes we may want to find the probability of more general sets. How would we do that? Well, for nice sets, you will just integrate the density over that nice set.

I'm not quite defining what "nice" means. That's a pretty technical topic in the theory of probability. But for our purposes, usually we will take b to be something like a union of intervals.

So how do you find the probability of falling in the union of two intervals? Well, you find the probability of falling in that interval plus the probability of falling in that interval. So it's the integral over this interval plus the integral over that interval.

And you think of this as just integrating over the union of the two intervals. So once you can calculate probabilities of intervals, then usually you are in business, and you can calculate anything else you might want. So the probability density function is a complete description of any statistical information we might be interested in for a continuous random variable.

OK. So now we can start walking through the concepts and the definitions that we have for discrete random variables and translate them to the continuous case. The first big concept is the concept of the expectation. One can start with a mathematical definition.

And here we put down a definition by just translating notation. Wherever we have a sum in the discrete case, we now write an integral. And wherever we had the probability mass function, we now throw in the probability density function.

This formula-- you may have seen it in freshman physics-- basically, it again gives you the center of gravity of the picture that you have when you have the density. It's the center of gravity of the object sitting underneath the probability density function. So that the interpretation still applies.

It's also true that our conceptual interpretation of what an expectation means is also valid in this case. That is, if you repeat an experiment a zillion times, each time drawing an independent sample of your random variable x , in the long run, the average that you are going to get should be the expectation. One can reason in a hand-waving way, sort of intuitively, the way we did it for the case of discrete random variables. But this is also a theorem of some sort. It's a limit theorem that we're going to visit later on in this class.

Having defined the expectation and having claimed that the interpretation of the expectation is that same as before, then we can start taking just any formula you've seen before and just translate it. So for example, to find the expected value of a function of a continuous random variable, you do not have to find the PDF or PMF of $g(X)$. You can just work directly with the original distribution of the random variable capital X .

And this formula is the same as for the discrete case. Sums get replaced by integrals. And PMFs get replaced by PDFs. And in particular, the variance of a random variable is defined again the same way.

The variance is the expected value, the average of the distance of X from the mean and then squared. So it's the expected value for a random variable that takes these numerical values. And same formula as before, integral and F instead of summation, and the P .

And the formulas that we have derived or formulas that you have seen for the discrete case, they all go through the continuous case. So for example, the useful relation for variances, which is this one, remains true. All right. So time for an example.

The most simple example of a continuous random variable that there is, is the so-called uniform random variable. So the uniform random variable is described by a density which is 0 except over an interval. And over that interval, it is constant. What is it meant to convey? It's trying to convey the idea that all x 's in this range are equally likely.

Well, that doesn't say very much. Any individual x has 0 probability. So it's conveying a little more than that. What it is saying is that if I take an interval of a given length δ , and I take another interval of the same length, δ , under the uniform distribution, these two intervals are going to have the same probability.

So being uniform means that intervals of same length have the same probability. So no interval is more likely than any other to occur. And in that sense, it conveys the idea of sort of complete randomness. Any little interval in our range is equally likely as any other little interval.

All right. So what's the formula for this density? I only told you the range. What's the height? Well, the area under the density must be equal to 1. Total probability is equal to 1. And so the

height, inescapably, is going to be 1 over (b minus a). That's the height that makes the density integrate to 1. So that's the formula.

And if you don't want to lose one point in your exam, you have to say that it's also 0, otherwise. OK. All right? That's sort of the complete answer.

How about the expected value of this random variable? OK. You can find the expected value in two different ways. One is to start with the definition. And so you integrate over the range of interest times the density. And you figure out what that integral is going to be.

Or you can be a little more clever. Since the center-of-gravity interpretation is still true, it must be the center of gravity of this picture. And the center of gravity is, of course, the midpoint. Whenever you have symmetry, the mean is always the midpoint of the diagram that gives you the PDF. OK. So that's the expected value of X .

Finally, regarding the variance, well, there you will have to do a little bit of calculus. We can write down the definition. So it's an integral instead of a sum. A typical value of the random variable minus the expected value, squared, times the density. And we integrate. You do this integral, and you find it's (b minus a) squared over that number, which happens to be 12.

Maybe more interesting is the standard deviation itself. And you see that the standard deviation is proportional to the width of that interval. This agrees with our intuition, that the standard deviation is meant to capture a sense of how spread out our distribution is. And the standard deviation has the same units as the random variable itself. So it's sort of good to-- you can interpret it in a reasonable way based on that picture.

OK, yes. Now, let's go up one level and think about the following. So we have formulas for the discrete case, formulas for the continuous case. So you can write them side by side. One has sums, the other has integrals.

Suppose you want to make an argument and say that something is true for every random variable. You would essentially need to do two separate proofs, for discrete and for continuous. Is there some way of dealing with random variables just one at a time, in one shot, using a sort of uniform notation? Is there a unifying concept?

Luckily, there is one. It's the notion of the cumulative distribution function of a random variable. And it's a concept that applies equally well to discrete and continuous random variables. So it's an object that we can use to describe distributions in both cases, using just one piece of notation.

So what's the definition? It's the probability that the random variable takes values less than a certain number little x . So you go to the diagram, and you see what's the probability that I'm falling to the left of this. And you specify those probabilities for all x 's.

In the continuous case, you calculate those probabilities using the integral formula. So you integrate from here up to x . In the discrete case, to find the probability to the left of some point, you go here, and you add probabilities again from the left.

So the way that the cumulative distribution function is calculated is a little different in the continuous and discrete case. In one case you integrate. In the other, you sum. But leaving aside how it's being calculated, what the concept is, it's the same concept in both cases. So let's see what the shape of the cumulative distribution function would be in the two cases.

So here what we want is to record for every little x the probability of falling to the left of x . So let's start here. Probability of falling to the left of here is 0-- 0, 0, 0. Once we get here and we start moving to the right, the probability of falling to the left of here is the area of this little rectangle. And the area of that little rectangle increases linearly as I keep moving. So accordingly, the CDF increases linearly until I get to that point.

At that point, what's the value of my CDF? 1. I have accumulated all the probability there is. I have integrated it.

This total area has to be equal to 1. So it reaches 1, and then there's no more probability to be accumulated. It just stays at 1. So the value here is equal to 1.

OK. How would you find the density if somebody gave you the CDF? The CDF is the integral of the density. Therefore, the density is the derivative of the CDF. So you look at this picture and take the derivative.

Derivative is 0 here, 0 here. And it's a constant up there, which corresponds to that constant. So more generally, and an important thing to know, is that the derivative of the CDF is equal to the density-- almost, with a little bit of an exception.

What's the exception? At those places where the CDF does not have a derivative-- here where it has a corner-- the derivative is undefined. And in some sense, the density is also ambiguous at that point. Is my density at the endpoint, is it 0 or is it 1?

It doesn't really matter. If you change the density at just a single point, it's not going to affect the value of any integral you ever calculate. So the value of the density at the endpoint, you can leave it as being ambiguous, or you can specify it.

It doesn't matter. So at all places where the CDF has a derivative, this will be true. At those places where you have corners, which do show up sometimes, well, you don't really care.

How about the discrete case? In the discrete case, the CDF has a more peculiar shape. So let's do the calculation. We want to find the probability of b to the left of here. That probability is 0, 0, 0. Once we cross that point, the probability of being to the left of here is 1/6. So as soon as we cross the point 1, we get the probability of 1/6, which means that the size of the jump that we have here is 1/6.

Now, question. At this point 1, which is the correct value of the CDF? Is it 0, or is it 1/6? It's 1/6 because-- you need to look carefully at the definitions, the probability of x being less than or equal to little x .

If I take little x to be 1, it's the probability that capital X is less than or equal to 1. So it includes the event that x is equal to 1. So it includes this probability here. So at jump points, the correct value of the CDF is going to be this one.

And now as I trace, x is going to the right. As soon as I cross this point, I have added another 3/6 probability. So that 3/6 causes a jump to the CDF. And that determines the new value. And finally, once I cross the last point, I get another jump of 2/6.

A general moral from these two examples and these pictures. CDFs are well defined in both cases. For the case of continuous random variables, the CDF will be a continuous function. It starts from 0. It eventually goes to 1 and goes smoothly-- well, continuously from smaller to higher values. It can only go up. It cannot go down since we're accumulating more and more probability as we are going to the right.

In the discrete case, again it starts from 0, and it goes to 1. But it does it in a staircase manner. And you get a jump at each place where the PMF assigns a positive mass. So jumps in the CDF are associated with point masses in our distribution. In the continuous case, we don't have any point masses, so we do not have any jumps either.

Now, besides saving us notation-- we don't have to deal with discrete and continuous twice-- CDFs give us actually a little more flexibility. Not all random variables are continuous or discrete. You can cook up random variables that are kind of neither or a mixture of the two.

An example would be, let's say you play a game. And with a certain probability, you get a certain number of dollars in your hands. So you flip a coin. And with probability 1/2, you get a reward of 1/2 dollars.

And with probability 1/2, you are led to a dark room where you spin a wheel of fortune. And that wheel of fortune gives you a random reward between 0 and 1. So any of these outcomes is possible. And the amount that you're going to get, let's say, is uniform.

So you flip a coin. And depending on the outcome of the coin, either you get a certain value or you get a value that ranges over a continuous interval. So what kind of random variable is it?

Is it continuous? Well, continuous random variables assign 0 probability to individual points. Is it the case here? No, because you have positive probability of obtaining 1/2 dollar. So our random variable is not continuous.

Is it discrete? It's not discrete, because our random variable can take values also over a continuous range. So we call such a random variable a mixed random variable.

If you were to draw its distribution very loosely, probably you would want to draw a picture like this one, which kind of conveys the idea of what's going on. So just think of this as a drawing of masses that are sitting over a table. We place an object that weighs half a pound, but it's an object that takes zero space. So half a pound is just sitting on top of that point. And we take another half-pound of probability and spread it uniformly over that interval.

So this is like a piece that comes from mass functions. And that's a piece that looks more like a density function. And we just throw them together in the picture. I'm not trying to associate any formal meaning with this picture. It's just a schematic of how probabilities are distributed, help us visualize what's going on.

Now, if you have taken classes on systems and all of that, you may have seen the concept of an impulse function. And you may start saying that, oh, I should treat this mathematically as a so-called impulse function. But we do not need this for our purposes in this class. Just think of this as a nice picture that conveys what's going on in this particular case.

So now, what would the CDF look like in this case? The CDF is always well defined, no matter what kind of random variable you have. So the fact that it's not continuous, it's not discrete shouldn't be a problem as long as we can calculate probabilities of this kind.

So the probability of falling to the left here is 0. Once I start crossing there, the probability of falling to the left of a point increases linearly with how far I have gone. So we get this linear increase. But as soon as I cross that point, I accumulate another $1/2$ unit of probability instantly. And once I accumulate that $1/2$ unit, it means that my CDF is going to have a jump of $1/2$.

And then afterwards, I still keep accumulating probability at a fixed rate, the rate being the density. And I keep accumulating, again, at a linear rate until I settle to 1. So this is a CDF that has certain pieces where it increases continuously. And that corresponds to the continuous part of our random variable. And it also has some places where it has discrete jumps. And those discrete jumps correspond to places in which we have placed a positive mass.

And by the-- OK, yeah. So this little 0 shouldn't be there. So let's cross it out. All right.

So finally, we're going to take the remaining time and introduce our new friend. It's going to be the Gaussian or normal distribution. So it's the most important distribution there is in all of probability theory. It plays a very central role. It shows up all over the place. We'll see later in the class in more detail why it shows up.

But the quick preview is the following. If you have a phenomenon in which you measure a certain quantity, but that quantity is made up of lots and lots of random contributions-- so your random variable is actually the sum of lots and lots of independent little random variables-- then invariability, no matter what kind of distribution the little random variables have, their sum will turn out to have approximately a normal distribution. So this makes the normal distribution to arise very naturally in lots and lots of contexts. Whenever you have noise that's comprised of lots of different independent pieces of noise, then the end result will be a random variable that's normal.

So we are going to come back to that topic later. But that's the preview comment, basically to argue that it's an important one. OK. And there's a special case. If you are dealing with a binomial distribution, which is the sum of lots of Bernoulli random variables, again you would expect that the binomial would start looking like a normal if you have many, many-- a large number of point fields.

All right. So what's the math involved here? Let's parse the formula for the density of the normal. What we start with is the function X squared over 2. And if you are to plot X squared over 2, it's a parabola, and it has this shape -- X squared over 2.

Then what do we do? We take the negative exponential of this. So when X squared over 2 is 0, then negative exponential is 1. When X squared over 2 increases, the negative exponential of that falls off, and it falls off pretty fast.

So as this goes up, the formula for the density goes down. And because exponentials are pretty strong in how quickly they fall off, this means that the tails of this distribution actually do go down pretty fast. OK. So that explains the shape of the normal PDF.

How about this factor 1 over square root 2 pi? Where does this come from? Well, the integral has to be equal to 1. So you have to go and do your calculus exercise and find the integral of this the minus X squared over 2 function and then figure out, what constant do I need to put in front so that the integral is equal to 1?

How do you evaluate that integral? Either you go to Mathematica or Wolfram's Alpha or whatever, and it tells you what it is. Or it's a very beautiful calculus exercise that you may have seen at some point. You throw in another exponential of this kind, you bring in polar coordinates, and somehow the answer comes beautifully out there.

But in any case, this is the constant that you need to make it integrate to 1 and to be a legitimate density. We call this the standard normal. And for the standard normal, what is the expected value? Well, the symmetry, so it's equal to 0.

What is the variance? Well, here there's no shortcut. You have to do another calculus exercise. And you find that the variance is equal to 1. OK. So this is a normal that's centered around 0.

How about other types of normals that are centered at different places? So we can do the same kind of thing. Instead of centering it at 0, we can take some place where we want to center it, write down a quadratic such as $(X - \mu)$ squared, and then take the negative exponential of that. And that gives us a normal density that's centered at μ .

Now, I may wish to control the width of my density. To control the width of my density, equivalently I can control the width of my parabola. If my parabola is narrower, if my parabola looks like this, what's going to happen to the density? It's going to fall off much faster.

OK. How do I make my parabola narrower or wider? I do it by putting in a constant down here. So by putting a σ here, this stretches or widens my parabola by a factor of σ . Let's see. Which way does it go?

If σ is very small, this is a big number. My parabola goes up quickly, which means my normal falls off very fast. So small σ corresponds to a narrower density.

And so it, therefore, should be intuitive that the standard deviation is proportional to sigma. Because that's the amount by which you are scaling the picture. And indeed, the standard deviation is sigma. And so the variance is sigma squared.

So all that we have done here to create a general normal with a given mean and variance is to take this picture, shift it in space so that the mean sits at mu instead of 0, and then scale it by a factor of sigma. This gives us a normal with a given mean and a given variance. And the formula for it is this one.

All right. Now, normal random variables have some wonderful properties. And one of them is that they behave nicely when you take linear functions of them. So let's fix some constants a and b, suppose that X is normal, and look at this linear function Y.

What is the expected value of Y? Here we don't need anything special. We know that the expected value of a linear function is the linear function of the expectation. So the expected value is this.

How about the variance? We know that the variance of a linear function doesn't care about the constant term. But the variance gets multiplied by a squared. So we get these variance, where sigma squared is the variance of the original normal.

So have we used so far the property that X is normal? No, we haven't. This calculation here is true in general when you take a linear function of a random variable. But if X is normal, we get the other additional fact that Y is also going to be normal. So that's the nontrivial part of the fact that I'm claiming here. So linear functions of normal random variables are themselves normal.

How do we convince ourselves about it? OK. It's something that we will do formerly in about two or three lectures from today. So we're going to prove it. But if you think about it intuitively, normal means this particular bell-shaped curve. And that bell-shaped curve could be sitting anywhere and could be scaled in any way.

So you start with a bell-shaped curve. If you take X, which is bell shaped, and you multiply it by a constant, what does that do? Multiplying by a constant is just like scaling the axis or changing the units with which you're measuring it. So it will take a bell shape and spread it or narrow it. But it will still be a bell shape. And then when you add the constant, you just take that bell and move it elsewhere.

So under linear transformations, bell shapes will remain bell shapes, just sitting at a different place and with a different width. And that sort of the intuition of why normals remain normals under this kind of transformation. So why is this useful?

Well, OK. We have a formula for the density. But usually we want to calculate probabilities. How will you calculate probabilities? If I ask you, what's the probability that the normal is less than 3, how do you find it? You need to integrate the density from minus infinity up to 3.

Unfortunately, the integral of the expression that shows up that you would have to calculate, an integral of this kind from, let's say, minus infinity to some number, is something that's not known in closed form. So if you're looking for a closed-form formula for this-- \bar{X} -- if you're looking for a closed-form formula that gives you the value of this integral as a function of \bar{X} , you're not going to find it. So what can we do?

Well, since it's a useful integral, we can just tabulate it. Calculate it once and for all, for all values of \bar{X} up to some precision, and have that table, and use it. That's what one does.

OK, but now there is a catch. Are we going to write down a table for every conceivable type of normal distribution-- that is, for every possible mean and every variance? I guess that would be a pretty long table. You don't want to do that.

Fortunately, it's enough to have a table with the numerical values only for the standard normal. And once you have those, you can use them in a clever way to calculate probabilities for the more general case. So let's see how this is done.

So our starting point is that someone has graciously calculated for us the values of the CDF, the cumulative distribution function, that is the probability of falling below a certain point for the standard normal and at various places. How do we read this table? The probability that X is less than, let's say, 0.63 is this number. This number, 0.7357, is the probability that the standard normal is below 0.63. So the table refers to the standard normal.

But someone, let's say, gives us some other numbers and tells us we're dealing with a normal with a certain mean and a certain variance. And we want to calculate the probability that the value of that random variable is less than or equal to 3. How are we going to do it? Well, there's a standard trick, which is so-called standardizing a random variable.

Standardizing a random variable stands for the following. You look at the random variable, and you subtract the mean. This makes it a random variable with 0 mean. And then if I divide by the standard deviation, what happens to the variance of this random variable? Dividing by a number divides the variance by sigma squared.

The original variance of X was σ^2 . So when I divide by σ , I end up with unit variance. So after I do this transformation, I get a random variable that has 0 mean and unit variance.

It is also normal. Why is its normal? Because this expression is a linear function of the X that I started with. It's a linear function of a normal random variable. Therefore, it is normal. And it is a standard normal.

So by taking a general normal random variable and doing this standardization, you end up with a standard normal to which you can then apply the table. Sometimes one calls this the normalized score. If you're thinking about test results, how would you interpret this number? It tells you how many standard deviations are you away from the mean.

This is how much you are away from the mean. And you count it in terms of how many standard deviations it is. So this number being equal to 3 tells you that X happens to be 3 standard deviations above the mean. And I guess if you're looking at your quiz scores, very often that's the kind of number that you think about. So it's a useful quantity.

But it's also useful for doing the calculation we're now going to do. So suppose that X has a mean of 2 and a variance of 16, so a standard deviation of 4. And we're going to calculate the probability of this event. This event is described in terms of this X that has ugly means and variances. But we can take this event and rewrite it as an equivalent event.

X less than 3 is this same as X minus 2 being less than 3 minus 2, which is the same as this ratio being less than that ratio. So I'm subtracting from both sides of the inequality the mean and then dividing by the standard deviation. This event is the same as that event.

Why do we like this better than that? We like it because this is the standardized, or normalized, version of X . We know that this is standard normal. And so we're asking the question, what's the probability that the standard normal is less than this number, which is $1/4$? So that's the key property, that this is normal $(0, 1)$.

And so we can look up now with the table and ask for the probability that the standard normal random variable is less than 0.25. Where is that going to be? 0.2, 0.25, it's here. So the answer is 0.987.

So I guess this is just a drill that you could learn in high school. You didn't have to come here to learn about it. But it's a drill that's very useful when we will be calculating normal probabilities all the time. So make sure you know how to use the table and how to massage a general normal random variable into a standard normal random variable.

OK. So just one more minute to look at the big picture and take stock of what we have done so far and where we're going. Chapter 2 was this part of the picture, where we dealt with discrete random variables. And this time, today, we started talking about continuous random variables. And we introduced the density function, which is the analog of the probability mass function.

We have the concepts of expectation and variance and CDF. And this kind of notation applies to both discrete and continuous cases. They are calculated the same way in both cases except that in the continuous case, you use sums. In the discrete case, you use integrals.

So on that side, you have integrals. In this case, you have sums. In this case, you always have F s in your formulas. In this case, you always have P s in your formulas.

So what's there that's left for us to do is to look at these two concepts, joint probability mass functions and conditional mass functions, and figure out what would be the equivalent concepts on the continuous side. So we will need some notion of a joint density when we're dealing with multiple random variables. And we will also need the concept of conditional density, again for the case of continuous random variables.

The intuition and the meaning of these objects is going to be exactly the same as here, only a little subtler because densities are not probabilities. They're rates at which probabilities accumulate. So that adds a little bit of potential confusion here, which, hopefully, we will fully resolve in the next couple of sections.

All right. Thank you.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013 Transcript – Recitation: Mean & Variance of the Exponential

Hi. In this video, we're going to compute some useful quantities for the exponential random variable.

So we're given that x is exponential with rate λ . PDF looks like this, and the formula is here.

First question, part a, what's the CDF? So let's go right in. The CDF of x is the probability that X is less than or equal to little x . Let's look at some cases here.

What if little x is less than 0? Well, x random variable only takes on these non-negative values. And so the probability that X is less than or equal to some negative number is going to be 0.

On the other hand, if x is greater than or equal to 0, we do actually have to integrate here. So to do that, we take the integral from minus infinity to x of $f_x(t)$ -- the dummy variable here used is t .

Notice that again, $f_x(t)$ is going to be 0 for negative values, so we take the integral here from 0. And now we plug in for $f_x(t)$. That's $-\lambda t e^{-\lambda t}$. And recall that the integral of u to the a is 1 over a times $e^{-\lambda a}$.

So here in this case, we'll get λ , which is just a constant. And then a here is going to be negative λ . So we get this, 0 to x . Lambdas cancel and we actually get 1 minus $e^{-\lambda x}$. So do this. And we are done with the CDF.

Now for the expectation. We use the standard formula, which is minus infinity to infinity $t f_x(t) dt$. So again, $f_x(t)$ is going to be 0 for a negative value. So we do the integral from 0. We get 0 to infinity $t \lambda e^{-\lambda t} dt$.

Now, you can try all you want to get rid of this t . It's not going to go even if you try all kinds of u substitution. But at the end the day, you're going to have to pull out your calculus textbook and find the integration by parts formula, which is $-v du$.

So the hope is that this integral is going to be easier than the one on the left. Notice that this is the integral of one of the terms here. And this is the derivative of one of the terms. So that may help you decide on how you select u and v .

In our case actually, I'm going to use u as t for u . Because when you take the derivative, it's going to become 1. And the derivative is what's going to go in that integral.

So this is going to be dt for du . And then, dv I'm going to select as whatever's left over. It's $\lambda e^{-\lambda t} dt$. So v is going to be-- we already did the integral-- $e^{-\lambda t}$.

And so if we do this, it's going to be negative t times $e^{-\lambda t}$ minus $\lambda e^{-\lambda t}$. So that's uv . Minus v , which is negative $e^{-\lambda t}$ times dt . That goes from 0 to infinity. This is evaluated from 0 to infinity.

Well, what does it mean for this to be evaluated from 0 to infinity? A better and easier way to look at this is to say, well, it's going to go from 0 to x . But then you take the limit as x goes to infinity. So that's going to help us here.

And this negative-- these negatives cancel. And we're left with-- let's plug in the bounds. We're left with negative x minus λx plus the integral of this is going to be 1 over negative $\lambda e^{-\lambda t}$ evaluated from 0 to infinity.

All right, so now the limit. So for the limit, notice that x increases as x goes to infinity. And this exponential decays. So they're kind of competing for each other. But the exponential is going to win because it decays way faster than x . And so this first term is going to go off-- the limit is going to go to 0.

All right. For this, if you evaluate the balance, the infinity makes this 0. And 0, you're going to get 1 over λ . So that's 1 over λ . All right. And so the expectation is 1 over λ .

OK, so now what's the variance? That's part c, right?

So we use the standard formula for variance, which is this. We already figured out the expectation. We just need to figure out the expectation of x^2 .

Well, we're just going to follow the same set of steps from before. For x^2 , it's just going to be t^2 , t^2 , t^2 , x^2 . The only thing that's going to change is what we choose for u here, for the u substitution. So it's going to be t^2 . So the derivative is going to change to $2t dt$. v is going to be exactly the same. And so here in this term, we get negative $2t e^{-\lambda t}$. But there's a negative sign out here, so the negatives cancel and we're left with a positive sign here. This is going to change. All right. OK.

So in order to do this integral, we can use a trick. We can move this-- so there's a $2t$ here. We move this 2 in here, leave the t inside. And you have to leave the t inside. But multiply by λ and divide by λ .

Now, look at that integral. 0 to infinity $t^2 \lambda e^{-\lambda t} dt$. Exactly the expectation that we computed. We already did that. That is just 1 over λ , so it's 2 over λ^2 .

Again, the limit as x goes to infinity-- the exponential will beat x^2 . No matter what polynomial we put in there, the exponential's going to win. So this is going to be 0 still. This

one's going to be 2 over lambda squared. So we're left with 2 over lambda squared for expectation of x^2 . And so we have 1 over lambda squared for the variance. OK, so we're done with the variance.

Part d. We're given that x_1 , x_2 , and x_3 are independent and identically distributed. They're exponentials with rate lambda. We're asked for the PDF of z , which is the max of x_1 , x_2 , and x_3 . How do we generally find a PDF?

We take the CDF and then take the derivative, right? We first find the CDF, and then take the derivative. So let's do that.

So first, let's see. Part d, find the CDF of z , which is going to be the probability that Z is less than or equal to little z , which is going to be equal to the probability that the max of x_1 , x_2 , x_3 is less than or equal to z . And this is going to have the same sort of structure as before.

If z is less than 0, x_1 , x_2 , x_3 are positive-- non-negative. And so this is the probability that if you get little z less than 0, you're not going to have any probability there. And so if z is greater than or equal to 0 is where it gets interesting. We need to do something special.

So the special thing here is to recognize that the probability of the max being less than or equal to z is actually also the probability of each of these random variables individually being less than or equal to z . Why is that true?

One way to check whether the events-- these two events are the same is to check the two directions. One direction say, if the max of x_1 , x_2 , x_3 is less than or equal to z , does that mean x_1 is less than or equal to z , x_2 is less than or equal to z , and x_3 is less than or equal to z ?

Yes. OK.

And then, if x_1 , x_2 , and x_3 are individually less than or equal to z , then the max is also less than or equal to z . So these two events are equivalent and this is true. By independence we can break this up. And we get-- these are all CDFs of the exponential and they all have this form. So it's just going to be 1 minus $e^{-\lambda z}$. Plug this in here. And then, try to take the derivative to get the PDF.

Let's see. So it's going to be the same, like this for z less than 0. For z greater than or equal to 0, it's going to be the derivative of this thing. Derivative of this thing is by chain rule, 3 times 1 minus $e^{-\lambda z}$. Then the derivative of negative $e^{-\lambda z}$, that's just $\lambda e^{-\lambda z}$. There we go. This is the PDF we were looking for.

So last problem. We're looking for the PDF of w , which is the min of x_1 and x_2 . So let's try this as a similar approach. Try the same thing, actually. See if it works.

So $w = \min(x_1, x_2)$. So let's see if this works.

Is it true that the min-- if the min of x_1 and x_2 is less than or equal to w , that each of them is less than or equal to w ? No, right?

X_1 could be less than or equal to w and x_2 could be bigger than w . And the min could still be less than or equal to w . So that's definitely not true. So what do we do here?

The trick is to flip it and say we want to compute the min of x_1 and x_2 being greater than w . In that case, let's check if we can do this trick.

If the min of x_1 and x_2 is greater than w , then clearly x_1 is bigger than w and x_2 is bigger than w . And if x_1 and x_2 are individually bigger than w , then clearly the min's also bigger than w . So this works. And now we can use independence as before.

And for this, this is just 1 minus the CDF here. So it's just going to be e to the minus lambda w for each of them. But that's the same as e to the minus lambda $2w$. Or e to the 2 lambda w . So it's going to be--

Notice the similarity between this and this. The only difference is this has a 2 lambda in there. That means that w is an exponential random variable with rate 2 lambda.

So then the PDF is going to be an exponential, whatever it is for an exponential. Except with rate 2 lambda.

You can also take the derivative of this and find that you get this. OK, so we're done with the problems. We computed some interesting quantities for the exponential random variable in this--

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013

Transcript – Tutorial: Normal Probability Calculation

Hi. In this video, we're going to do standard probability calculations for normal random variables. We're given that x is standard normal with mean 0 and variance 1. And y is normal with mean one and variance 4. And we're asked for a couple of probabilities.

For the normal CDF, we don't have a closed form expression. And so people generally tabulate values and for the standard normal case. So if we want little x equal to 3.49, we just look for 3.4 along the rows and 0.09 along the columns, and then pick the value appropriately. So for part A, we're asked what's the probability that x is less than equal to 1.5?

That's exactly ϕ of 1.5 and we can look that up. 1.5 directly and that's 0.9332. Then we're asked, what's the probability that x is less than equal to negative 1? Notice that negative values are not on this table. And the reason that is is because the standard normal is symmetric around zero. And we don't really need that.

We just recognize that the area in this region is exactly the area in this region. And so that's equal to the probability that x is greater than equal to 1. This is equal to 1 minus the probability that x is less than 1. And we can put the equal sign in here because x is continuous, it doesn't matter.

And so we're going to get, this is equal to 1 minus ϕ of 1, which is 1.00, and that's 0.8413. OK.

For part B, we're asked for this distribution of y minus 1 over 2. So any linear function of a normal random variable is also normal. And you can see that by using the derived distribution for linear functions of random variables.

So in this case, we only need to figure out what's the mean and the variance of this normal random variable. So the mean in this case, I'm going to write that as y over 2 minus 1/2. The expectation operator is linear and so that's going to be-- and the expectation in this case is 1, so that's going to be 0.

Now the variance. For the shift, it doesn't affect the spread. And so the variance is exactly going to be the same without the minus 1/2. And for the constant, you can just pull that out and square it. And the variance of y we know is 4. And so that's 1/4 times 4, that's 1.

OK. So now we know that y minus 1 over 2 is actually standard normal. Actually for any normal random variable, you can follow the same procedure. You just subtract its mean, which is 1 in this case. And divide by its standard deviation and you will get a standard normal distribution.

All right, so for part C we want the probability that y is between negative 1 and 1. So let's try to massage it so that we can use the standard normal table. And we already know that this is standard normal, so let's subtract both sides by negative 1. And that's equal to-- I'm going to call

this standard normal z , so that's easier to write. And that's equal to negative 1 less than equal to z , less than equal to zero. So we're looking for this region, 0, 1, negative 1. So that's just the probability that it's less than zero minus probability that it's less than negative 1.

Well for a standard normal, half the mass is below zero and a half the mass is above. And so that's just going to be 0.5 directly. And for this, we've already computed this for a standard normal, which was x in our case. And that was 1 minus 0.8413. Done.

So we basically calculated a few standard probabilities for normal distributions. And we did that by looking them up from the standard normal table.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

LECTURE 8

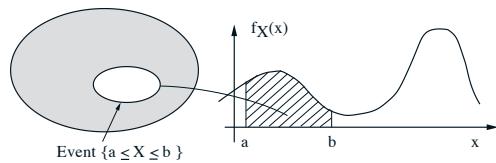
- **Readings:** Sections 3.1-3.3

Lecture outline

- Probability density functions
- Cumulative distribution functions
- Normal random variables

Continuous r.v.'s and pdf's

- A continuous r.v. is described by a probability density function f_X



$$\mathbf{P}(a \leq X \leq b) = \int_a^b f_X(x) dx$$

$$\int_{-\infty}^{\infty} f_X(x) dx = 1$$

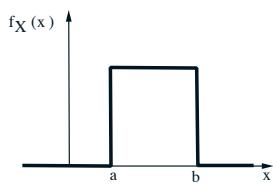
$$\mathbf{P}(x \leq X \leq x + \delta) = \int_x^{x+\delta} f_X(s) ds \approx f_X(x) \cdot \delta$$

$$\mathbf{P}(X \in B) = \int_B f_X(x) dx, \quad \text{for "nice" sets } B$$

Means and variances

- $\mathbf{E}[X] = \int_{-\infty}^{\infty} x f_X(x) dx$
- $\mathbf{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx$
- $\text{var}(X) = \sigma_X^2 = \int_{-\infty}^{\infty} (x - \mathbf{E}[X])^2 f_X(x) dx$

- **Continuous Uniform r.v.**



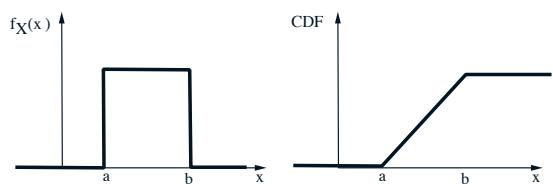
- $f_X(x) = c \quad a \leq x \leq b$

- $\mathbf{E}[X] =$

- $\sigma_X^2 = \int_a^b \left(x - \frac{a+b}{2} \right)^2 \frac{1}{b-a} dx = \frac{(b-a)^2}{12}$

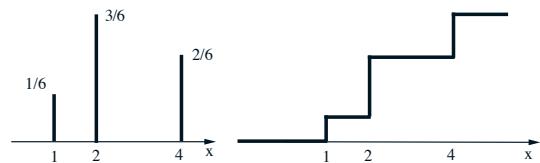
Cumulative distribution function (CDF)

$$F_X(x) = \mathbf{P}(X \leq x) = \int_{-\infty}^x f_X(t) dt$$



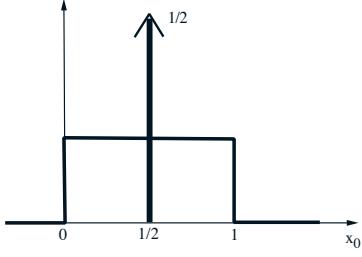
- Also for discrete r.v.'s:

$$F_X(x) = \mathbf{P}(X \leq x) = \sum_{k \leq x} p_X(k)$$



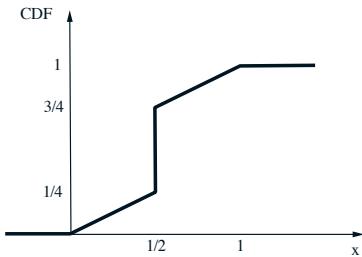
Mixed distributions

- Schematic drawing of a combination of a PDF and a PMF



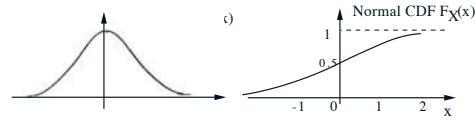
- The corresponding CDF:

$$F_X(x) = P(X \leq x)$$



Gaussian (normal) PDF

- Standard normal $N(0, 1)$: $f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$



- $E[X] =$ $\text{var}(X) = 1$

- General normal $N(\mu, \sigma^2)$:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

- It turns out that:
 $E[X] = \mu$ and $\text{Var}(X) = \sigma^2$.

- Let $Y = aX + b$

- Then: $E[Y] =$ $\text{Var}(Y) =$

- Fact: $Y \sim N(a\mu + b, a^2\sigma^2)$

Calculating normal probabilities

- No closed form available for CDF
 - but there are tables
(for standard normal)
- If $X \sim N(\mu, \sigma^2)$, then $\frac{X - \mu}{\sigma} \sim N(0, 1)$
- If $X \sim N(2, 16)$:

$$P(X \leq 3) = P\left(\frac{X - 2}{4} \leq \frac{3 - 2}{4}\right) = \text{CDF}(0.25)$$

	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817

The constellation of concepts

$p_X(x)$	$f_X(x)$
$F_X(x)$	
$E[X], \text{ var}(X)$	
$p_{X,Y}(x, y)$	$f_{X,Y}(x, y)$
$p_{X Y}(x y)$	$f_{X Y}(x y)$

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Recitation 8
October 5, 2010

1. Let Z be a continuous random variable with probability density function

$$f_z(z) = \begin{cases} \gamma(1 + z^2), & \text{if } -2 < z < 1, \\ 0, & \text{otherwise.} \end{cases}$$

- (a) For what value of γ is this possible?
- (b) Find the cumulative distribution function of Z .

2. Problem 3.9, pages 186–187 in the text.

The taxi stand and the bus stop near Al's home are in the same location. Al goes there at a given time and if a taxi is waiting, (this happens with probability $2/3$) he boards it. Otherwise he waits for a taxi or a bus to come, whichever comes first. The next taxi will arrive in a time that is uniformly distributed between 0 and 10 minutes, while the next bus will arrive in exactly 5 minutes. Find the CDF and the expected value of Al's waiting time.

3. Let λ be a positive number. The continuous random variable X is called **exponential** with parameter λ when its probability density function is

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

- (a) Find the cumulative distribution function (CDF) of X .
- (b) Find the mean of X .
- (c) Find the variance of X .
- (d) Suppose X_1 , X_2 , and X_3 are independent exponential random variables, each with parameter λ . Find the PDF of $Z = \max\{X_1, X_2, X_3\}$.
- (e) Find the PDF of $W = \min\{X_1, X_2\}$.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Recitation 8 Solutions
October 5, 2010

1. (a) We know that the PDF must integrate to 1. Therefore we have

$$\int_{-\infty}^{\infty} f_Z(z) dz = \int_{-2}^1 \gamma(1+z^2) = \gamma \left(z + \frac{1}{3}z^3 \right) \Big|_{-2}^1 = 6\gamma.$$

From this we conclude $\gamma = 1/6$.

- (b) To find the CDF, we integrate:

$$\begin{aligned} F_Z(z) &= \int_{-\infty}^z f_Z(t) dt = \begin{cases} 0, & \text{if } z < -2, \\ \frac{1}{6} \left(t + \frac{1}{3}t^3 \right) \Big|_{-2}^z, & \text{if } -2 \leq z \leq 1, \\ 1, & \text{if } z > 1 \end{cases} \\ &= \begin{cases} 0, & \text{if } z < -2, \\ \frac{1}{6} \left(z + \frac{1}{3}z^3 + \frac{14}{3} \right), & \text{if } -2 \leq z \leq 1, \\ 1, & \text{if } z > 1. \end{cases} \end{aligned}$$

2. See textbook, Problem 3.9, page 187.

3. (a) For $x \geq 0$,

$$F_X(x) = \int_{-\infty}^x f_X(t) dt = \int_0^x \lambda e^{-\lambda t} dt = [-e^{-\lambda t}]_0^x = 1 - e^{-\lambda x}.$$

For $x < 0$, we have $F_X(x) = \int_{-\infty}^x f_X(t) dt = 0$. Thus we conclude

$$F_X(x) = \begin{cases} 0, & \text{if } x < 0, \\ 1 - e^{-\lambda x}, & \text{if } x \geq 0. \end{cases}$$

- (b) The key step in the following computation uses integration by parts, whereby

$$\int_0^\infty u dv = uv \Big|_0^\infty - \int_0^\infty v du$$

is applied with $u = x$ and $v = -e^{-\lambda x}$:

$$\mathbf{E}[X] = \int_{-\infty}^\infty x f_X(x) dx = \int_0^\infty x \lambda e^{-\lambda x} dx = [-xe^{-\lambda x}]_0^\infty + \int_0^\infty e^{-\lambda x} dx = \frac{1}{\lambda}.$$

- (c) Integrating by parts with $u = x^2$ and $v = -e^{-\lambda x}$ in the second line below gives

$$\begin{aligned} \mathbf{E}[X^2] &= \int_{-\infty}^\infty x^2 f_X(x) dx = \int_0^\infty x^2 \lambda e^{-\lambda x} dx \\ &= [-x^2 e^{-\lambda x}]_0^\infty + 2 \int_0^\infty x e^{-\lambda x} dx = \frac{2}{\lambda} \mathbf{E}[X] = \frac{2}{\lambda^2}. \end{aligned}$$

Combining with the previous computation, we obtain

$$\text{var}(X) = \mathbf{E}[X^2] - (\mathbf{E}[X])^2 = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda} \right)^2 = \frac{1}{\lambda^2}.$$

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
 (Fall 2010)

- (d) The maximum of a set is upper bounded by z when each element of the set is upper bounded by z . Thus for any positive z ,

$$\begin{aligned}\mathbf{P}(Z \leq z) &= \mathbf{P}(\max\{X_1, X_2, X_3\} \leq z) = \mathbf{P}(X_1 \leq z, X_2 \leq z, X_3 \leq z) \\ &= \mathbf{P}(X_1 \leq z)\mathbf{P}(X_2 \leq z)\mathbf{P}(X_3 \leq z) \\ &= (1 - e^{-\lambda z})^3,\end{aligned}$$

where the third equality uses the independence of X_1 , X_2 , and X_3 . Thus,

$$F_Z(z) = \begin{cases} 0, & \text{if } z < 0, \\ (1 - e^{-\lambda z})^3, & \text{if } z \geq 0. \end{cases}$$

Differentiating the CDF gives the desired PDF:

$$f_Z(z) = \begin{cases} 0, & \text{if } z < 0, \\ 3\lambda e^{-\lambda z}(1 - e^{-\lambda z})^2, & \text{if } z \geq 0. \end{cases}$$

- (e) The minimum of a set is lower bounded by w when each element of the set is lower bounded by w . Thus for any positive w ,

$$\begin{aligned}\mathbf{P}(W \geq w) &= \mathbf{P}(\min\{X_1, X_2\} \geq w) = \mathbf{P}(X_1 \geq w, X_2 \geq w) \\ &= \mathbf{P}(X_1 \leq w)\mathbf{P}(X_2 \leq w) \\ &= (e^{-\lambda w})^2 = e^{-2\lambda w}\end{aligned}$$

where the third equality uses the independence of X_1 and X_2 . Thus,

$$F_W(w) = \begin{cases} 0, & \text{if } w < 0, \\ 1 - e^{-2\lambda w}, & \text{if } w \geq 0. \end{cases}$$

We can recognize this as the CDF of an exponential random variable with parameter 2λ . The PDF is

$$f_W(w) = \begin{cases} 0, & \text{if } w < 0, \\ 2\lambda e^{-2\lambda w}, & \text{if } w \geq 0. \end{cases}$$

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Tutorial 4
October 7/8, 2010

1. Let X and Y be Gaussian random variables, with $X \sim N(0, 1)$ and $Y \sim N(1, 4)$.

- Find $\mathbf{P}(X \leq 1.5)$ and $\mathbf{P}(X \leq -1)$.
- What is the distribution of $\frac{Y-1}{2}$?
- Find $\mathbf{P}(-1 \leq Y \leq 1)$.

2. Example 3.15, page 169 in text.

Ben throws a dart at a circular target of radius r . We assume that he always hits the target, and that all points of impact (x, y) are equally likely. Compute the joint PDF $f_{X,Y}(x, y)$ of the random variables X and Y and compute the conditional PDF $f_{X|Y}(x|y)$.

3. Problem 3.20, page 191 in text.

An absent-minded professor schedules two student appointments for the same time. The appointment durations are independent and exponentially distributed with mean thirty minutes. The first student arrives on time, but the second student arrives five minutes late. What is the expected value of the time between the arrival of the first student and the departure of the second student?

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2010

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Tutorial 4: Solutions

1. (a)

$$\begin{aligned}\mathbf{P}(X \leq 1.5) &= \Phi(1.5) \\ &= 0.9332.\end{aligned}$$

$$\begin{aligned}\mathbf{P}(X \leq -1) &= 1 - \mathbf{P}(X \leq 1) \\ &= 1 - \Phi(1) \\ &= 1 - 0.8413 \\ &= 0.1587.\end{aligned}$$

(b)

$$\begin{aligned}\mathbf{E}\left[\frac{Y-1}{2}\right] &= \frac{1}{2}(\mathbf{E}[Y] - 1) \\ &= 0.\end{aligned}$$

$$\begin{aligned}\text{var}\left(\frac{Y-1}{2}\right) &= \text{var}\left(\frac{Y}{2}\right) \\ &= \frac{1}{4}\text{var}Y \\ &= 1.\end{aligned}$$

Thus, the distribution of $\frac{Y-1}{2}$ is $\mathcal{N}(0, 1)$.

(c)

$$\begin{aligned}\mathbf{P}(-1 \leq Y \leq 1) &= \mathbf{P}\left(\frac{-1-1}{2} \leq \frac{Y-1}{2} \leq \frac{1-1}{2}\right) \\ &= \Phi(0) - \Phi(-1) \\ &= \Phi(0) - (1 - \Phi(1)) \\ &= 0.3413.\end{aligned}$$

2. Example 3.15, page 169 in text. See solutions in the text.

3. Problem 3.20, page 191 in text. See online solutions.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013 Transcript – Lecture 9

The following content is provided under a Creative Commons license. Your support will help MIT OpenCourseWare continue to offer high-quality educational resources for free. To make a donation or view additional materials from hundreds of MIT courses, visit MIT OpenCourseWare at ocw.mit.edu.

JOHN TSITSIKLIS: OK let's start. So we've had the quiz. And I guess there's both good and bad news in it. Yesterday, as you know, the bad news. The average was a little lower than what we would have wanted. On the other hand, the good news is that the distribution was nicely spread. And that's the main purpose of this quiz is basically for you to calibrate and see roughly where you are standing.

The other piece of the good news is that, as you know, this quiz doesn't count for very much in your final grade. So it's really a matter of calibration and to get your mind set appropriately to prepare for the second quiz, which counts a lot more. And it's more substantial. And we'll make sure that the second quiz will have a higher average.

All right. So let's go to our material. We're talking now these days about continuous random variables. And I'll remind you what we discussed last time. I'll remind you of the concept of the probability density function of a single random variable. And then we're going to rush through all the concepts that we covered for the case of discrete random variables and discuss their analogs for the continuous case. And talk about notions such as conditioning independence and so on.

So the big picture is here. We have all those concepts that we developed for the case of discrete random variables. And now we will just talk about their analogs in the continuous case. We already discussed this analog last week, the density of a single random variable.

Then there are certain concepts that show up both in the discrete and the continuous case. So we have the cumulative distribution function, which is a description of the probability distribution of a random variable and which applies whether you have a discrete or continuous random variable. Then there's the notion of the expected value.

And in the two cases, the expected value is calculated in a slightly different way, but not very different. We have sums in one case, integrals in the other. And this is the general pattern that we're going to have. Formulas for the discrete case translate to corresponding formulas or expressions in the continuous case. We generically replace sums by integrals, and we replace must functions with density functions.

Then the new pieces for today are going to be mostly the notion of a joint density function, which is how we describe the probability distribution of two random variables that are somehow related, in general, and then the notion of a conditional density function that tells us the distribution of one random variable X when you're told the value of another random variable Y . There's another concept, which is the conditional PDF given that the certain event has happened.

This is a concept that's in some ways simpler. You've already seen a little bit of that in last week's recitation and tutorial.

The idea is that we have a single random variable. It's described by a density. Then you're told that the certain event has occurred. Your model changes the universe that you are dealing with. In the new universe, you are dealing with a new density function, the one that applies given the knowledge that we have that the certain event has occurred.

All right. So what exactly did we say about continuous random variables? The first thing is the definition, that a random variable is said to be continuous if we are given a certain object that we call the probability density function and we can calculate interval probabilities given this density function. So the definition is that the random variable is continuous if you can calculate probabilities associated with that random variable given that formula. So this formula tells you that the probability that your random variable falls inside this interval is the area under the density curve.

OK. There's a few properties that a density function must satisfy. Since we're talking about probabilities, and probabilities are non-negative, we have that the density function is always a non-negative function. The total probability over the entire real line must be equal to 1. So the integral when you integrate over the entire real line has to be equal to 1. That's the second property.

Another property that you get is that if you let a equal to b , this integral becomes 0. And that tells you that the probability of a single point in the continuous case is always equal to 0. So these are formal properties.

When you want to think intuitively, the best way to think about what the density function is to think in terms of little intervals, the probability that my random variable falls inside the little interval. Well, inside that little interval, the density function here is roughly constant. So that integral becomes the value of the density times the length of the interval over which you are integrating, which is Δ .

And so the density function basically gives us probabilities of little events, of small events. And the density is to be interpreted as probability per unit length at a certain place in the diagram. So in that place in the diagram, the probability per unit length around this neighborhood would be the height of the density function at that point.

What else? We have a formula for calculating expected values of functions of random variables. In the discrete case, we had the formula where here we had the sum, and instead of the density, we had the PMF. The same formula is also valid in the continuous case. And it's not too hard to derive, but we will not do it.

But let's think of the intuition of what this formula says. You're trying to figure out on the average how much $g(X)$ is going to be. And then you reason, and you say, well, X may turn out to take a particular value or a small interval of values. This is the probability that X falls inside the small interval. And when that happens, $g(X)$ takes that value.

So this fraction of the time, you fall in the little neighborhood of x , and you get so much. Then you average over all the possible x 's that can happen. And that gives you the average value of the function $g(X)$.

OK. So this is the easy stuff. Now let's get to the new material. We want to talk about multiple random variables simultaneously. So we want to talk now about two random variables that are continuous, and in some sense that they are jointly continuous.

And let's see what this means. The definition is similar to the definition we had for a single random variable, where I take this formula here as the definition of continuous random variables. Two random variables are said to be jointly continuous if we can calculate probabilities by integrating a certain function that we call the joint density function over the set of interest.

So we have our two-dimensional plane. This is the x - y plane. There's a certain event S that we're interested in. We want to calculate the probability. How do we do that?

We are given this function $f_{(X,Y)}$, the joint density. It's a function of the two arguments x and y . So think of that function as being some kind of surface that sits on top of the two-dimensional plane. The probability of falling inside the set S , we calculate it by looking at the volume under the surface, that volume that sits on top of S . So the surface underneath it has a certain total volume.

What should that total volume be? Well, we think of these volumes as probabilities. So the total probability should be equal to 1. The total volume under this surface, should be equal to 1. So that's one property that we want our density function to have. So when you integrate over the entire space, this is of the volume under your surface. That should be equal to 1. Of course, since we're talking about probabilities, the joint density should be a non-negative function.

So think of the situation as having one pound of probability that's spread all over your space. And the height of this joint density function basically tells you how much probability tends to be accumulated in certain regions of space as opposed to other parts of the space. So wherever the density is big, that means that this is an area of the two-dimensional plane that's more likely to occur. Where the density is small, that means that those x - y 's are less likely to occur.

You have already seen one example of continuous densities. That was the example we had in the very beginning of the class with a uniform distribution on the unit square. That was a special case of a density function that was constant. So all places in the unit square were roughly equally likely as any other places.

But in other models, some parts of the space may be more likely than others. And we describe those relative likelihoods using this density function. So if somebody gives us the density function, this determines for us probabilities of all the subsets of the two-dimensional plane.

Now for an intuitive interpretation, it's good to think about small events. So let's take a particular x here and then x plus delta. So this is a small interval. Take another small interval here that goes from y to y plus delta. And let's look at the event that x falls here and y falls right there.

What is this event? Well, this is the event that will fall inside this little rectangle. Using this rule for calculating probabilities, what is the probability of that rectangle going to be? Well, it should be the integral of the density over this rectangle. Or it's the volume under the surface that sits on top of that rectangle.

Now, if the rectangle is very small, the joint density is not going to change very much in that neighborhood. So we can treat it as a constant. So the volume is going to be the height times the area of the base. The height at that point is whatever the function happens to be around that point. And the area of the base is delta squared.

So this is the intuitive way to understand what a joint density function really tells you. It specifies for you probabilities of little squares, of little rectangles. And it allows you to think of the joint density function as probability per unit area. So these are the units of the density, its probability per unit area in the neighborhood of a certain point.

So what do we do with this density function once we have it in our hands? Well, we can use it to calculate expected values. Suppose that you have a function of two random variables described by a joint density. You can find, perhaps, the distribution of this random variable and then use the basic definition of the expectation. Or you can calculate expectations directly, using the distribution of the original random variables.

This is a formula that's again identical to the formula that we had for the discrete case. In the discrete case, we had a double sum here, and we had PMFs. So the intuition behind this formula is the same that one had for the discrete case. It's just that the mechanics are different.

Then something that we did in the discrete case was to find a way to go from the joint density of the two random variables taken together to the density of just one of the random variables. So we had a formula for the discrete case. Let's see how things are going to work out in the continuous case.

So in the continuous case, we have here our two random variables. And we have a density for them. And let's say that we want to calculate the probability that x falls inside this interval. So we're looking at the probability that our random variable X falls in the interval from little x to x plus delta.

Now, by the properties that we already have for interpreting the density function of a single random variable, the probability of a little interval is approximately the density of that single random variable times delta. And now we want to find a formula for this marginal density in terms of the joint density.

OK. So this is the probability that x falls inside this interval. In terms of the two-dimensional plane, this is the probability that (x,y) falls inside this strip. So to find that probability, we need to calculate the probability that (x,y) falls in here, which is going to be the double integral over the interval over this strip, of the joint density.

And what are we integrating over? y goes from minus infinity to plus infinity. And the dummy variable x goes from little x to x plus delta. So to integrate over this strip, what we do is for any given y , we integrate in this dimension. This is the x integral. And then we integrate over the y dimension.

Now what is this inner integral? Because x only varies very little, this is approximately constant in that range. So the integral with respect to x just becomes delta times $f(x,y)$. And then we've got our dy .

So this is what the inner integral will evaluate to. We are integrating over the little interval. So we're keeping y fixed. Integrating over here, we take the value of the density times how much we're integrating over. And we get this formula.

OK. Now, this expression must be equal to that expression. So if we cancel the deltas, we see that the marginal density must be equal to the integral of the joint density, where we have integrated out the value of y . So this formula should come as no surprise at this point.

It's exactly the same as the formula that we had for discrete random variables. But now we are replacing the sum with an integral. And instead of using the joint PMF, we are using the joint PDF.

Then, continuing going down the list of things we did for discrete random variables, we can now introduce a definition of the notion of independence of two random variables. And by analogy with the discrete case, we define independence to be the following condition. Two random variables are independent if and only if their joint density function factors out as a product of their marginal densities. And this property needs to be true for all x and y . So this is the formal definition.

Operationally and intuitively, what does it mean? Well, intuitively it means the same thing as in the discrete case. Knowing anything about X shouldn't tell you anything about Y . That is, information about X is not going to change your beliefs about Y . We are going to come back to this statement in a second.

The other thing that it allows you to do-- I'm not going to derive this-- is it allows you to calculate probabilities by multiplying individual probabilities. So if you ask for the probability that x falls in a certain set A and y falls in a certain set B , then you can calculate that probability by multiplying individual probabilities. This takes just two lines of derivation, which I'm not going to do. But it comes back to the usual notion of independence of events. Basically, operationally independence means that you can multiply probabilities.

So now let's look at an example. There's a sort of pretty famous and classical one. It goes back a lot more than a 100 years. And it's the famous Needle of Buffon. Buffon was a French naturalist who, for some reason, also decided to play with probability. And look at the following problem.

So you have the two-dimensional plane. And on the plane we draw a bunch of parallel lines. And those parallel lines are separated by a length. And the lines are apart at distance d . And we throw

a needle at random, completely at random. And we'll have to give a meaning to what "completely at random" means.

And when we throw a needle, there's two possibilities. Either the needle is going to fall in a way that does not intersect any of the lines, or it's going to fall in a way that it intersects one of the lines. We're taking the needle to be shorter than this distance, so the needle cannot intersect two lines simultaneously. It either intersects 0, or it intersects one of the lines. The question is to find the probability that the needle is going to intersect a line.

What's the probability of this? OK. We are going to approach this problem by using our standard four-step procedure. Set up your sample space, describe a probability law on that sample space, identify the event of interest, and then calculate. These four steps basically correspond to these three bullets and then the last equation down here.

So first thing is to set up a sample space. We need some variables to describe what happened in the experiment. So what happens in the experiment is that the needle lands somewhere. And where it lands, we can describe this by specifying the location of the center of the needle.

And what do we mean by the location of the center? Well, we can take as our variable to be the distance from the center of the needle to the nearest line. So it tells us the vertical distance of the center of the needle from the nearest line.

The other thing that matters is the orientation of the needle. So we need one more variable, which we take to be the angle that the needle is forming with the lines. We can put the angle here, or you can put in there. Yes, it's still the same angle.

So we have these two variables that described what happened in the experiment. And we can take our sample space to be the set of all possible x 's and θ 's. What are the possible x 's? The lines are d apart, so the nearest line is going to be anywhere between 0 and $d/2$ away. So that tells us what the possible x 's will be.

As for θ , it really depends how you define your angle. We are going to define our θ to be the acute angle that's formed between the needle and a line, if you were to extend it. So θ is going to be something between 0 and $\pi/2$. So I guess these red pieces really correspond to the part of setting up the sample space. OK. So that's part one.

Second part is we need a model. OK. Let's take our model to be that we basically know nothing about how the needle falls. It can fall in any possible way, and all possible ways are equally likely.

Now, if you have those parallel lines, and you close your eyes completely and throw a needle completely at random, any x should be equally likely. So we describe that situation by saying that X should have a uniform distribution. That is, it should have a constant density over the range of interest.

Similarly, if you kind of spin your needle completely at random, any angle should be as likely as any other angle. And we decide to model this situation by saying that theta also has a uniform distribution over the range of interest. And finally, where we put it should have nothing to do with how much we rotate it. And we capture this mathematically by saying that X is going to be independent of theta.

Now, this is going to be our model. I'm not deriving the model from anything. I'm only saying that this sounds like a model that does not assume any knowledge or preference for certain values of x rather than other values of theta. In the absence of any other particular information you might have in your hands, that's the most reasonable model to come up with. So you model the problem that way.

So what's the formula for the joint density? It's going to be the product of the densities of X and Theta. Why is it the product? This is because we assumed independence.

And the density of X , since it's uniform, and since it needs to integrate to 1, that density needs to be $2/d$. That's the density of X . And the density of Theta needs to be $2/\pi$. That's the value for the density of Theta so that the overall probability over this interval ends up being 1.

So now we do have our joint density in our hands. The next thing to do is to identify the event of interest. And this is best done in a picture. And there's two possible situations that one could have. Either the needle falls this way, or it falls this way.

So how can we tell if one or the other is going to happen? It has to do with whether this interval here is smaller than that or bigger than that. So we are comparing the height of this interval to that interval. This interval here is capital X .

This interval here, what is it? This is half of the length of the needle, which is $l/2$. To find this height, we take $l/2$ and multiply it with the sine of the angle that we have. So the length of this interval up here is $l/2 \times \sin \theta$.

If this is smaller than x , the needle does not intersect the line. If this is bigger than x , then the needle intersects the line. So the event of interest, that the needle intersects the line, is described this way in terms of x and theta.

And now that we have the event of interest described mathematically, all that we need to do is to find the probability of this event, we integrate the joint density over the part of (x, θ) space in which this inequality is true. So it's a double integral over the set of all x 's and θ 's where this is true. The way to do this integral is we fix θ , and we integrate for x 's that go from 0 up to that number. And θ can be anything between 0 and $\pi/2$. So the integral over this set is basically this double integral here.

We already have a formula for the joint density. It's $4/\pi d$, so we put it here. And now, fortunately, this is a pretty easy integral to evaluate. The integral with respect to x -- there's nothing in here. So the integral is just the length of the interval over which we're integrating. It's $l/2 \sin \theta$.

And then we need to integrate this with respect to theta. We know that the integral of a sine is a negative cosine. You plug in the values for the negative cosine at the two end points. I'm sure you can do this integral . And we finally obtain the answer, which is amazingly simple for such a pretty complicated-looking problem. It's $2l$ over πd .

So some people a long, long time ago, after they looked at this answer, they said that maybe that gives us an interesting way where one could estimate the value by π , for example, experimentally. How do you do that? Fix l and d , the dimensions of the problem. Throw a million needles on your piece of paper. See how often your needless do intersect the line.

That gives you a number for this quantity. You know l and d , so you can use that to infer π . And there's an apocryphal story about a wounded soldier in a hospital after the American Civil War who actually had heard about this and was spending his time in the hospital throwing needles on pieces of paper. I don't know if it's true or not. But let's do something similar here.

So let's look at this diagram. We fix the dimensions. This is supposed to be our little d . That's supposed to be our little l . We have the formula from the previous slide that p is $2l$ over πd . In this instance, we choose d to be twice l . So this number is $1/\pi$. So the probability that the needle hits the line is $1/\pi$.

So I need needles that are 3.1 centimeters long. I couldn't find such needles. But I could find paper clips that are 3.1 centimeters long. So let's start throwing paper clips at random and see how many of them will end up intersecting the lines. Good.

OK. So out of eight paper clips, we have exactly four that intersected the line. So our estimate for the probability of intersecting the line is $1/2$, which gives us an estimate for the value of π , which is two. Well, I mean, within an engineering approximation, we're in the right ballpark, right?

So this might look like a silly way of trying to estimate π . And it probably is. On the other hand, this kind of methodology is being used especially by physicists and also by statisticians. It's used a lot. When is it used?

If you have an integral to calculate, such as this integral, but you're not lucky, and your functions are not so simple where you can do your calculations by hand, and maybe the dimensions are larger-- instead of two random variables you have 100 random variables, so it's a 100-fold integral-- then there's no way to do that in the computer. But the way that you can actually do it is by generating random samples of your random variables, doing that simulation over and over many times. That is, by interpreting an integral as a probability, you can use simulation to estimate that probability. And that gives you a way of calculating integrals.

And physicists do actually use that a lot, as well as statisticians, computer scientists, and so on. It's a so-called Monte Carlo method for evaluating integrals. And it's a basic piece of the toolbox in science these days.

Finally, the harder concept of the day is the idea of conditioning. And here things become a little subtle when you deal with continuous random variables. OK. First, remember again our basic interpretation of what a density is. A density gives us probabilities of little intervals.

So how should we define conditional densities? Conditional densities should again give us probabilities of little intervals, but inside a conditional world where we have been told something about the other random variable. So what we would like to be true is the following. We would like to define a concept of a conditional density of a random variable X given the value of another random variable Y . And it should behave the following way, that the conditional density gives us the probability of little intervals-- same as here-- given that we are told the value of y .

And here's where the subtleties come. The main thing to notice is that here I didn't write "equal," I wrote "approximately equal." Why do we need that?

Well, the thing is that conditional probabilities are not defined when you condition on an event that has 0 probability. So we need the conditioning event here to have posed this probability. So instead of saying that Y is exactly equal to little y , we want to instead say we're in a new universe where capital Y is very close to little y .

And then this notion of "very close" kind of takes the limit and takes it to be infinitesimally close. So this is the way to interpret conditional probabilities. That's what they should mean.

Now, in practice, when you actually use probability, you forget about that subtlety. And you say, well, I've been told that Y is equal to 1.3. Give me the conditional distribution of X . But formally or rigorously, you should say I'm being told that Y is infinitesimally close to 1.3. Tell me the distribution of X .

Now, if this is what we want, what should this quantity be? It's a conditional probability, so it should be the probability of two things happening-- X being close to little x , Y being close to little y . And that's basically given to us by the joint density divided by the probability of the conditioning event, which has something to do with the density of Y itself. And if you do things carefully, you see that the only way to satisfy this relation is to define the conditional density by this particular formula.

OK. Big discussion to come down in the end to what you should have probably guessed by now. We just take any formulas and expressions from the discrete case and replace PMFs by PDFs. So the conditional PDF is defined by this formula where here we have joint PDF and marginal PDF, as opposed to the discrete case where we had the joint PMF and the marginal PMF. So in some sense, it's just a syntactic change.

In another sense, it's a little subtler on how you actually interpret it. Speaking about interpretation, what are some ways of thinking about the joint density? Well, the best way to think about it is that somebody has fixed little y for you. So little y is being fixed here. And we look at this density as a function of X .

I've told you what Y is. Tell me what you know about X. And you tell me that X has a certain distribution. What does that distribution look like? It has exactly the same shape as the joint density.

Remember, we fixed Y. So this is a constant. So the only thing that varies is X. So we get the function that behaves like the joint density when you fix y, which is really you take the joint density, and you take a slice of it.

You fix a y, and you see how it varies with x. So in that sense, the conditional PDF is just a slice of the joint PDF. But we need to divide by a certain number, which just scales it and changes its shape.

We're coming back to a picture in a second. But before going to the picture, let's go back to the interpretation of independence. If the two random variables are independent, according to our definition in the previous slide, the joint density is going to factor as the product of the marginal densities. The density of Y in the numerator cancels the density in the denominator. And we're just left with the density of X.

So in the case of independence, what we get is that the conditional is the same as the marginal. And that solidifies our intuition that in the case of independence, being told something about the value of Y does not change our beliefs about how X is distributed. So whatever we expected about X is going to remain true even after we are told something about Y.

So let's look at some pictures. Here is what the joint PDF might look like. Here we've got our x and y-axis. And if you want to calculate the probability of a certain event, what you do is you look at that event and you see how much of that mass is sitting on top of that event.

Now let's start slicing. Let's fix a value of x and look along that slice where we obtain this function. Now what does that slice do? That slice tells us for that particular x what the possible values of y are going to be and how likely they are.

If we integrate over all y's, what do we get? Integrating over all y's just gives us the marginal density of X. It's the calculation that we did here.

By integrating over all y's, we find the marginal density of X. So the total area under that slice gives us the marginal density of X. And by looking at the different slices, we find how likely the different values of x are going to be.

How about the conditional? If we're interested in the conditional of Y given X, how would you think about it? This refers to a universe where we are told that capital X takes on a specific value.

So we put ourselves in the universe where this line has happened. There's still possible values of y that can happen. And this shape kind of tells us the relative likelihoods of the different y's. And this is indeed going to be the shape of the conditional distribution of Y given that X has occurred.

On the other hand, the conditional distribution must add up to 1. So the total probability over all of the different y 's in this universe, that total probability should be equal to 1. Here it's not equal to 1. The total area is the marginal density. To make it equal to 1, we need to divide by the marginal density, which is basically to renormalize this shape so that the total area under that slice, under that shape, is equal to 1.

So we start with the joint. We take the slices. And then we adjust the slices so that every slice has an area underneath equal to 1. And this gives us the conditional.

So for example, down here-- you can not even see it in this diagram-- but after you renormalize it so that its total area is equal to 1, you get this sort of narrow spike that goes up. And so this is a plot of the conditional distributions that you get for the different values of x . Given a particular value of x , you're going to get this certain conditional distribution.

So this picture is worth about as much as anything else in this particular chapter. Make sure you kind of understand exactly all these pieces of the picture. And finally, let's go, in the remaining time, through an example where we're going to throw in the bucket all the concepts and notations that we have introduced so far. So the example is as follows.

We start with a stick that has a certain length. And we break it at a completely random location. And-- yes, this 1 should be l . OK. So it has length l . And we're going to break it at the random place. And we call that random place where we break it, we call it X .

X can be anywhere, uniform distribution. So this means that X has a density that goes from 0 to l . I guess this capital L is supposed to be the same as the lower-case l . So that's the density of X . And since the density needs to integrate to 1, the height of that density has to be $1/l$.

Now, having broken the stick and given that we are left with this piece of the stick, I'm now going to break it again at a completely random place, meaning I'm going to choose a point where I break it uniformly over the length of the stick. What does this mean? And let's call Y the location where I break it. So Y is going to range between 0 and x . x is the stick that I'm left with.

So I'm going to break it somewhere in between. So I pick a y between 0 and x . And of course, x is less than l . And I'm going to break it there. So y is uniform between 0 and x .

What does that mean, that the density of y , given that you have already told me x , ranges from 0 to little x ? If I told you that the first break happened at a particular x , then y can only range over this interval. And I'm assuming a uniform distribution over that interval. So we have this kind of shape. And that fixes for us the height of the conditional density.

So what's the joint density of those two random variables? By the definition of conditional densities, the conditional was defined as the ratio of this divided by that. So we can find the joint density by taking the marginal and then multiplying by the conditional. This is the same formula as in the discrete case. This is our very familiar multiplication rule, but adjusted to the case of continuous random variables. So P s become F s.

OK. So we do have a formula for this. What is it? It's $1/x$ -- that's the density of X -- times $1/x$, which is the conditional density of Y . This is the formula for the joint density.

But we must be careful. This is a formula that's not valid anywhere. It's only valid for the x 's and y 's that are possible.

And the x 's and y 's that are possible are given by these inequalities. So x can range from 0 to 1, and y can only be smaller than x . So this is the formula for the density on this part of our space. The density is 0 anywhere else.

So what does it look like? It's basically a $1/x$ function. So it's sort of constant along that dimension. But as x goes to 0, your density goes up and can even blow up. It sort of looks like a sail that's raised and somewhat curved and has a point up there going to infinity. So this is the joint density.

Now once you have in your hands a joint density, then you can answer in principle any problem. It's just a matter of plugging in and doing computations. How about calculating something like a conditional expectation of Y given a value of x ? OK. That's a concept we have not defined so far. But how should we define it?

Means the reasonable thing. We'll define it the same way as ordinary expectations except that since we're given some conditioning information, we should use the probability distribution that applies to that particular situation. So in a situation where we are told the value of x , the distribution that applies is the conditional distribution of Y . So it's going to be the conditional density of Y given the value of x .

Now, we know what this is. It's given by $1/x$. So we need to integrate y times $1/x$ dy.

And what should we integrate over? Well, given the value of x , y can only range from 0 to x . So this is what we get. And you do your integral, and you get that this is $x/2$.

Is it a surprise? It shouldn't be. This is just the expected value of Y in a universe where X has been realized and Y is given by this distribution. Y is uniform between 0 and x . The expected value of Y should be the midpoint of this interval, which is $x/2$.

Now let's do fancier stuff. Since we have the joint distribution, we should be able to calculate the marginal. What is the distribution of Y ? After breaking the stick twice, how big is the little piece that I'm left with? How do we find this?

To find the marginal, we just take the joint and integrate out the variable that we don't want. A particular y can happen in many ways. It can happen together with any x . So we consider all the possible x 's that can go together with this y and average over all those x 's.

So we plug in the formula for the joint density from the previous slide. We know that it's $1/x$. And what's the range of the x 's? So to find the density of Y for a particular y up here, I'm going

to integrate over x 's. The density is 0 here and there. The density is nonzero only in this part. So I need to integrate over x 's going from here to there.

So what's the "here"? This line goes up at the slope of 1. So this is the line x equals y . So if I fix y , it means that my integral starts from a value of x that is also equal to y .

So where the integral starts from is at x equals y . And it goes all the way until the end of the length of our stick, which is 1. So we need to integrate from little y up to 1.

So that's something that almost always comes up. It's not enough to have just this formula for integrating the joint density. You need to keep track of different regions. And if the joint density is 0 in some regions, then you exclude those regions from the range of integration. So the range of integration is only over those values where the particular formula is valid, the places where the joint density is nonzero.

All right. The integral of $1/x \, dx$, that gives you a logarithm. So we evaluate this integral, and we get an expression of this kind. So the density of Y has a somewhat unexpected shape. So it's a logarithmic function.

And it goes this way. It's for y going all the way to 1. When y is equal to 1, the logarithm of 1 is equal to 0. But when y approaches 0, logarithm of something big blows up, and we get a shape of this form.

OK. Finally, we can calculate the expected value of Y . And we can do this by using the definition of the expectation. So integral of y times the density of y . We already found what that density is, so we can plug it in here. And we're integrating over the range of possible y 's, from 0 to 1.

Now this involves the integral for $y \log y$, which I'm sure you have encountered in your calculus classes but maybe do not remember how to do it. In any case, you look it up in some integral tables or do it by parts. And you get the final answer of $1/4$.

And at this point, you say, that's a really simple answer. Shouldn't I have expected it to be $1/4$? I guess, yes. I mean, when you break it once, the expected value of what you are left with is going to be $1/2$ of what you started with. When you break it the next time, the expected length of what you're left with should be $1/2$ of the piece that you are now breaking.

So each time that you break it at random, you expected it to become smaller by a factor of $1/2$. So if you break it twice, you are left something that's expected to be $1/4$. This is reasoning on the average, which happens to give you the right answer in this case. But again, there's the warning that reasoning on the average doesn't always give you the right answer. So be careful about doing arguments of this type.

Very good. See you on Wednesday.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013 Transcript – Recitation: Probability that Three Pieces Form a Triangle

In this problem, we're going to look at the probability that when you take a stick and break it into three pieces randomly that these three pieces can actually be used to form a triangle. All right, so we start out with a stick of unit length, so-- length 1. And we'll choose a point along the stick to break. And we'll choose that point uniformly at random.

So let's say that we chose it here, that was the point where we'll break it. And then independently of this first choice we'll again choose a second point to break it. Again, uniformly at random along the entire stick. So let's say the second point we chose was here. So what we have now is, we'll break it here, here, and so we'll have three pieces-- the first one, the left one and the right one.

And we want to know, what's the probability that when you take these three pieces you could form a triangle? So the first thing we should ask ourselves is, what do you need-- what conditions must be satisfied in order to actually be able to form a triangle with three pieces? So you could think about, what would stop you from being able to do that?

Well, one possibility is that you have pieces that look like this. So in that case you would try to form something that looks like this. But you can't get a triangle because these two pieces are too short and they can't touch each other. So actually the condition that must be satisfied is that when you take any two of the three pieces, their combined length has to be greater than the length of the remaining third piece. And that has to be true for any two pieces.

And really that's just so that any two pieces, they can touch and still form a triangle. So let's try to add some probability to this. So we have a unit length stick. So let's actually give a coordinate system. The stick goes from 0 to 1. And let's say that we break it at these two points.

So the first point where we choose, we'll call that x . So that's the first point that we choose to break it. And then the second point we choose, we'll call that y . Now note that I've drawn it so that x is to the left of y . But it could actually be the case that the first point I chose is here and the second point that I chose is to the left.

But for now, let's first assume that this scenario holds. That the first point is to the left of the second point. So under this assumption, we can see that-- from the definition of these random variables-- we can actually see that the lengths are given by these three lengths.

So the lengths are x , the left most piece has length x . The second, middle piece has length y minus x . And the last piece has length 1 minus y . And now let's recall our three conditions.

So the conditions were that any two of these, the sum of any two lengths, has to be at least-- has to be greater than the length of the third piece. So let's do these together. So x plus y minus x has to be greater than 1 minus y . So with these two pieces you can cover this third piece.

We also need that with the first and third pieces, we can cover the middle piece. And we need with the second and third pieces, we can cover the first piece. Now this looks kind of messy, but in fact we can actually simplify this.

So this actually simplifies. $x - x$, that disappears. And so this actually simplifies to $2y$ has to be at least 1. Or even more simply, y has to be greater than $1/2$.

What about this one? This one, we can rearrange things again. x we can move over. y we can move over here. And we get that $2x + 1$ has to be greater than $2y$. Or put in other words, y is less than $x + 1/2$.

And for the last one, again we can simplify. The y 's cancel each other out. And we're left with $2x$ is less than 1. Or x is less than $1/2$. So these are our three conditions that need to be satisfied.

So now we just have to figure out what's the probability that this is actually satisfied? Now let's go back to original definition and see what are the actual distributions for these random variables, x and y . Remember, we defined them to be x is the location of the first break and y is the location of the second break. And as we said in the problem, these are chosen uniformly at random and they're independent.

And so we can actually draw out their joint PDF. So x and y , you can cover any point in the square. And moreover, it's actually uniform within the square. Because each one is chosen uniformly at random and they're independent. So it's anywhere in here.

And so what do we need to do? We just need to identify, what is the probability that these three conditions hold? Rewrite this, line these up. So these are our three conditions that we need. And now remember, we're still working under the assumption that the first point that we chose is actually to the left of the second point.

So what does that mean? That means that we are actually in this top triangle, top half-- x is less than y . All right, so what do we need? We need y to be at least $1/2$, so here's $1/2$. So we need y to be above this line.

We need x to be less than $1/2$. So we need x to be to the left of here. So now so far we're stuck in this upper square. And the last thing we need is y to be less than $x + 1/2$.

What is y ? The line y equals x and a $1/2$, $x + 1/2$, is this one. So y has to be less than that, so it would have to be in this triangle here. So these three conditions tell us that in order for us to have a triangle we need to for x and y to fall jointly in this small triangle here.

Now because the joint distribution is uniform, we know that the density is just 1, right? Because the area here is just 1. So the height is just 1 as well. And so the density, or the probability of falling within this small triangle, is just going to be also the area of this triangle.

And what is the area of this triangle? Well, you can fit 8 of these triangles in here, or you could think of it as $1/2$ times $1/2$ times $1/2$. So the area is just $1/8$. So assuming that x is less than y , then the probability of forming a triangle is $1/8$.

Now, that's only half this story, though. Because it's possible that when you chose these two break points that we actually had the opposite result. That x , the point that you chose first, falls to the right of the point that you chose second.

In which case everything kind of flips. Now we assume that y is less than x , which means that now we're in this lower triangle in the square. Now we can go through this whole exercise again.

But really, what we can see is that all we've really done is just swap the names. Instead of having x and y we now call x -- we call x y and we call y x . And so if we just swap names, we can see that-- let's just fast forward through all these steps and see that we could just swap names here, too, as well, in the three conditions.

So instead of needing y to be greater than $1/2$, we just need x to be greater than $1/2$. Instead of having x less than $1/2$, we need y less than $1/2$. We also swap this.

So we need x to be less than y plus $1/2$ or y is greater than x minus $1/2$. All right, now let's figure out what this corresponds to. We need x to be greater than $1/2$, so it needs to be to the right of here. We need y to be less than $1/2$, so we need it to be below this line. And we need y to be greater than x minus $1/2$.

What is the line y equals x minus $1/2$? That is this line here. And we need y to be greater than that, so it needs to be above this line. And so we get that this is the triangle, the small triangle that we need in this case. And notice that it's exactly the same area as this one, right? And so we get another contribution of $1/8$ here.

So the final answer is $1/8$ plus $1/8$ is $1/4$. So the probability of forming a triangle using this three pieces is exactly $1/4$. And so notice that we've done is, you've set things up very methodically in the beginning by assigning these random variables. And you consider different cases.

Because you don't actually know the order in which x and y might fall, let's just assume that one particular order and work from there. And then do the other case, as well. And it just so happened that because of the symmetry of the problem the second case was actually very simple. We could just see that it is actually symmetric and so we get the same answer.

So this is kind of an interesting problem because it's actually a practical application of something that you might actually do. And you can see that just by applying these probability concepts you can actually--

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013

Transcript – Tutorial: The Absent Minded Professor

Hi. In this problem, we have an absent-minded professor who will inadvertently give us some practice with exponential random variables. So the professor has made two appointments with two students and inadvertently made them at the same time. And what we do is we model the duration of these appointments with an exponential random variable.

So remember, an exponential random variable is a continuous random variable that takes on non-negative values, and it's parametrized by a rate parameter, lambda. And the exponential random variable is often used to model durations of time-- so time until something happens, so for example, in this case, time until the student leaves or the appointment is over. Or sometimes you will also use it to be as a model of time until something fails.

And one thing that will be useful is the CDF of this exponential random variable. So the probability that it's less than or equal to some value, little t, is equal to 1 minus e to the minus lambda t. So this is, of course, valid only when t is non-negative.

The other useful property is that the expected value of an exponential random variable is just 1 over the parameter lambda. And the last thing that we'll use specifically in this problem is the memory list property of exponential random variables. And so recall that that just means that if you pop in the middle of an exponential random variable, the distribution for that random variable going forward from the point where you popped in is exactly the same as if it had just started over. So that's why we call it the memory list property. Basically, the past doesn't really matter, and going forward from whatever point that you observe it at, it looks as if it had just started over afresh.

And last thing we'll use, which is a review of a concept from earlier, is total expectation. So let's actually model this problem with two random variables. Let's let T1 be the time that the first student takes in the appointment and T2 be the time that the second student takes. And what we're told in the problem is that they're both exponential with mean 30 minutes.

So remember the mean being 30 minutes means that the lambda is 1 over the mean. And so the lambda in this case would be 1/30. And importantly, we're also told that they are independent. So how long the first person takes is independent of how long the second person takes.

So the first student arrives on time and take some random amount of time, T1. The second student arrives exactly five minutes late. And whatever the second person meets with the professor, that student will then take some random amount of time, T2. What we're asked to do is find the expected time between when the first student arrives-- so we can just call that time 0-- and when the second student leaves.

Now you may say, well we're dealing with expectations, so it's easier. And in this case, it probably is just the expectation of how long the first student takes plus the expectation of how

long the second student takes. So it should be about 60 minutes or exactly 60 minutes. Now, why is that not exactly right? It's because there is a small wrinkle, that the students may not go exactly back to back.

So let's actually draw out a time frame of what might actually happen. So here's time 0, when the first student arrives. And the first will go for some amount of time and leave.

And now let's consider two scenarios. One scenario is that the first student takes more than five minutes to complete. Well then the second student will have arrived at 5 minutes and then will be already waiting whenever this first student leaves. So then the second student will immediately pick up and continue. And in that case, we do have two exponentials back to back.

But there could be another situation. Suppose that the first student didn't take very long at all and finished within five minutes, in which case the second student hasn't arrived yet. So this professor is idle in between here. And so we actually don't necessarily have two of them going back to back. So there's an empty period in between that we have to account for.

So with that in mind, we see that we have two scenarios. And so what does that beg to use? Well, we can split them up into the two scenarios and then calculate expectations with each one and then use total expectation to find the overall expected length of time.

OK, so let's begin with the first scenario. The first scenario is that, let's say, the first student finished within five minutes. So what does that mean in terms of the definitions that we've used? That means T_1 is less than or equal to 5.

So if the first student took less than five minutes, then what happens? Then we know that the amount of time that you'd need to take-- let's call that something else. Let's call that X . So X is the random variable that we're interested in, the time between when the first student comes and the second student leaves. This is the value that we want to find.

Well we know that we're guaranteed that there will be a five-minute interval. So first student will come, and then the second person will take over. So we're guaranteed that the first five minutes will be the difference between when time starts and when the second student arrives.

And then, after that, it's just however long the second student takes, which is just the expected value of T_2 . And T_2 is an exponential random variable with mean 30. So in this case, it's just 35.

So the first student doesn't take very long. Then we just get the five minutes, that little buffer, plus however long the second student takes, which, on average, is 30 minutes. Now what is the probability of this happening?

The probability of this happening is the probability that the first student takes less than five minutes. And here is where we use the CDF that we wrote out earlier. It's going to be $1 - e^{-\lambda t}$. So in this case, t is five and λ is $1/30$. So it's $1 - e^{-5/30}$ is the probability.

All right, now let's consider the second case. The second case is that the first student actually takes longer than five minutes. OK, so what happens in that case?

Here's five minutes. The first student came to five minutes. The second student arrived, and the first student is still going. So he goes for some amount of time. And then whenever he finishes, the second student continues. So now the question is, what is the total amount of time in this case?

Well, you can think of it as using the memory list property. This is where it comes in. So the first five minutes, we know that it was already taken because we're considering the second scenario, which we're given that T_1 is greater than 5.

And so the question now is, if we know that, how much longer does it take? How much longer past the five-minute mark does the first student take? And by the member list property, we know that it's as if the first student started over. So there was no memory of the first five minutes, and it's as if the first student just arrived also at the five-minute mark and met with the professor. So past the five-minute mark, it's as if you have a new exponential random variable, still with mean 30.

And so what we get is that, in this case, you get the guaranteed five minutes, and then you get the memory list continuation of the first student's appointment. So you get another 30 minutes on average because of the memory list property. And then whenever the first student finally does finish up, the second student will immediately take over because he has already arrived. It's past the five-minute mark. And then that second student will take, again, on average, 30 more minutes.

So what you get is, in this case, the appointment lasts 65 minutes on average. Now what is the probability of this case? The probability of this case is the probability that T_1 is greater than 5. And now we know that that is just the complement of this, 1 minus that. So it's just e to the minus $5/30$.

So now we have both scenarios. We have the probabilities of each scenario, and we have the expectation under each scenario. Now all that remains now is to combine them using total expectation.

So I really should have written expectation of X given T_1 is less than or equal to 5 here. And this is expectation of X given that T_1 is greater than 5. So expectation of X overall is the probability that T_1 is less than or equal to 5 times the expectation of X given that T_1 is less than or equal to 5 plus the probability that T_1 is greater than 5 times the expectation of X given that T_1 is greater than 5. And we have all four of these pieces here. so it's 35 times 1 minus e to the minus $5/30$ plus 65 times e to the minus $5/30$. And it turns out that this is approximately equal to 60.394 minutes.

All right, so what have we found? We found that the original guess that we had, if we just had two meetings back to back, was on average it would take 60 minutes. It turns out that, because of

the way that things are set up, because of the five minute thing, it actually takes a little longer than 60 minutes on average.

And why is that? It's because the five sometimes adds an extra buffer, adds a little bit of extra amount, because it would have been shorter in this scenario because, if the both students had arrived on time, then the second student would have been able to pick up right here immediately. And so both appointments would have ended sooner. But because the second student didn't arrive until five minutes, there was some empty space that was wasted. And that's where you get you the little bit of extra time.

So this is a nice problem just to get some more exercise with exponential random variables and also nicely illustrates the memory list property, which was a key points in order to solve this. And it also is nice because we get to review a useful tool that we've been using all course long, which is to split things into different scenarios and then solve the simpler problems and then combine them up, for example using total expectation. So I hope that was helpful, and see you next time.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013

Transcript – Recitation: Uniform Probabilities on a Triangle

Hi. In this problem, we're going to get a bunch of practice working with multiple random variables together. And so we'll look at joint PDFs, marginal PDFs, conditional PDFs, and also get some practice calculating expectations as well. So the problem gives us a pair of random variables-- x and y . And we're told that the joint distribution is uniformly distributed on this triangle here, with the vertices being 0, 0 1, 0, and 0, 1. So it's uniform in this triangle.

And the first part of the problem is just to figure out what exactly is disjoint PDF of the two random variables. So in this case, it's pretty easy to calculate, because we have a uniform distribution. And remember, when you have a uniform distribution, you can just imagine it being a sort of plateau coming out of the board. And it's flat.

And so the height of the plateau, in order to calculate it, you just need to figure out what the area of this thing is, of this triangle is. So remember, when you had single random variables, what we had to do was calculate, for uniform distribution, we had to integrate to 1. So you took the length, and you took 1 over the length was the correct scaling factor.

Here, you take the area. And the height has to make it so that the entire volume here integrates to 1. So the joint PDF is just going to be 1 over whatever this area is. And the area is pretty simple to calculate. It's $1/2$ base times height. So it's $1/2$.

And so what we have is that the area is $1/2$. And so the joint PDF of x and y is going to equal 2. But remember, you always have to be careful when writing these things to remember the ranges when these things are valid. So it's only 2 within this triangle.

And outside of the triangle, it's 0. So what exactly does inside the triangle mean? Well, we can write it more mathematically. So this diagonal line, it's given by x plus y equals 1. So everything in the triangle is really x plus y is less than or equal to 1. It means everything under this triangle.

And so we need x plus y to be less than or equal to 1 and also x to be non-negative and y to be non-negative. So with these inequalities, that captures everything within this triangle. And otherwise, the joint PDF is going to be 0.

The next part asks us to find, using this joint PDF, the marginal of y . And remember, when you have a joint PDF of two random variables, you essentially have everything that you need, because from this joint PDF, you can calculate marginals, you can calculate from the margins, you can calculate conditionals. The joint PDF captures everything that there is to know about this pair of random variables.

Now, to calculate a marginal PDF of y , remember a marginal really just means collapsing the other random variable down. And so you can just imagine taking this thing and collapsing it

down onto the y-axis. And mathematically, that is just saying that we integrate out the other random variable.

So the other random variable in this case will be x. We take x and we get rid of it by integrating out from negative infinity to infinity. Of course, this joint PDF is 0 in a lot of places. And so a lot of these will be 0. And only for a certain range of x's will this integral actually be non-zero.

And so again, the other time when we have to be careful is when we have these limits of integration, we need to make sure that we have the right limits. And so we know that the joint PDF is 2. It's nonzero only within this triangle. And so it's only 2 within this triangle, which means what for x?

Well, depending on what x and y are, this will be either 2 or 0. So let's just fix some value of y. Pretend that we've picked some value y, let's say here. We want this value of y.

Well, what are the values of x such that the joint PDF for that value y is actually nonzero, it's actually 2? Well, it's everything from x equals 0 to whatever x value this is. But this x value, actually, if you think about it, is just 1 minus y, because this line is x plus y equals 1. So whatever y is, x is going to be 1 minus that.

And so the correct limits would actually be from 0 to 1 minus y. And then the rest of that is pretty simple. You integrate this. This is a pretty simple integral. And you get that it's actually two times 1 minus y. That's a y.

But of course, again, we need to make sure that we have the right regions. So this is not always true for y, of course. This is only true for y between 0 and 1. And otherwise, it's actually 0, because when you take a y down here, well, there's no values of x that will give you a nonzero joint PDF.

And if you take a value of y higher than this, the same thing happens. So we can actually draw this out and see what it looks like. So let's actually draw a small picture here. Here's y.

Here's the marginal PDF of y. And here's 2. And it actually looks like this. It's a triangle and a 0 outside this range.

So does that make sense? Well, first of all, you see that actually does in fact integrates to 1, which is good. And the other thing we notice is that there is a higher density for smaller values of y. So why is that?

Why are smaller values of y more likely than larger values of y? Well, because when you have smaller values of y, you're down here. And it's more likely because there are more values of x that go along with it that make that value of y more likely to appear.

Say you have a large value of y. Then you're up here at the tip. Well, there aren't very many combinations of x and y that give you that large a value of y. And so that large value of y becomes less likely.

Another way to think about it is, when you collapse this down, there's a lot more stuff to collapse down its base. There's a lot of x 's to collapse down. But up here, there's only a very little bit of x to collapse down. And the PDF of y becomes more skewed towards smaller values of y .

So now, the next thing that we want to do is calculate the conditional PDF of x , given y . Well, let's just recall what that means. This is what we're looking for-- the conditional PDF of x , given y .

And remember, this is calculated by taking the joint and dividing by the marginal of y . So we actually have the top and the bottom. We have to joint PDF from part A. And from part B, we calculated the marginal PDF of y . So we have both pieces.

So let's actually plug them in. Again, the thing that you have to be careful here is about the ranges of x and y where these things are valid, because this is only non-zero when x and y fall within this triangle. And this is only non-zero when y is between 0 and 1. So we need to be careful.

So the top, when it's non-zero, it's 2. And the bottom, when it's non-zero, it's 2 times 1 minus y . So we can simplify that to be 1 over 1 minus y . And when is this true?

Well, it's true when x and y are in the triangle and y is between 0 and 1. So put another way, that means that this is valid when y is between 0 and 1 and x is between 0 and 1 minus y , because whatever x has to be, it has to be such that they actually still fall within this triangle. And outside of this, it's 0.

So let's see what this actually looks like. So this is x , and this is the conditional PDF of x , given y . Let's say this is 1 right here.

Then what it's saying is, let's say we're given that y is some little y . Let's say it's somewhere here. Then it's saying that the conditional PDF of x given y is this thing. But notice that this value, 1 over 1 minus y , does not depend on x . So in fact, it actually is uniform.

So it's uniform between 0 and 1 minus y . And the height is something like 1 over 1 minus y . And this is so that the scaling makes it so that actually is a valid PDF, because the integral is to 1.

So why is the case? Why is that when you condition on y being some value, you get that the PDF of x is actually uniform? Well, when you look over here, let's again just pretend that you're taking this value of y .

Well, when you're conditioning on y being this value, you're basically taking a slice of this joint PDF at this point. But remember, the original joint PDF was uniform. So when you take a slice of a uniform distribution, joint uniform distribution, you still get something that is uniform. Just imagine that you have a cake that is flat.

Now, you take a slice at this level. Then whatever slice you have is also going to be imagine being a flat rectangle. So it's still going to be uniform. And that's why the conditional PDF of x given y is also uniform.

Part D now asks us to find a conditional expectation of x . So we want to find the expectation of x , given that y is some little y . And for this, we can use the definition. Remember, expectations are really just weighted sums. Or in the [? continuous ?] case, it's an integral.

So you take the value. And then you weight it by the density. And in this case, because we're taking conditional a expectation, what we weight it by is the conditional density. So it's the conditional density of x given that y is little y . We integrate with respect to x .

And fortunately, we know what this conditional PDF is, because we calculated it earlier in part C. And we know that it's this-- 1 over 1 minus y . But again, we have to be careful, because this formula, 1 over 1 minus y , is only valid certain cases. So let's think about this first.

Let's think about some extreme cases. What if y , little y , is negative? If little y is negative, we're conditioning on something over here. And so there is no density for y being negative or for y , say, in other cases when y is greater than 1.

And so in those cases, this expectation is just undefined, because conditioning on that doesn't really make sense, because there's no density for those values of y . Now, let's consider the case that actually makes, sense where y is between 0 and 1. Now, we're in business, because that is the range where this formula is valid.

So this formula is valid, and we can plug it in. So it's 1 over 1 minus y dx . And then the final thing that we again need to check is what the limits of this integration is. So we're integrating with respect to x . So we need to write down what values of x , what ranges of x is this conditional PDF valid.

Well, luckily, we specified that here. x has to be between 0 and 1 minus y . So let's actually calculate this integral. This 1 over 1 minus y is a constant with respect to x . You can just pull that out.

And then now, you're really just integrating x from 0 to 1 minus y . So the integral of x is [? 1 ?], $1/2x$ squared. So you get a $1/2x$ squared, and you integrate that from 0 to 1 minus y . And so when you plug in the limits, you'll get a 1 minus y squared. That will cancel out the 1 over 1 minus y .

And what you're left with is just 1 minus y over 2. And again, we have to specify that this is only true for y between 0 and 1. Now, we can again actually verify that this makes sense. What we're really looking for is the conditional expectation of x given some value of y .

And we already said that condition on y being some value of x is uniformly distributed between 0 and 1 minus y . And so remember for our uniform distribution, the expectation is simple. It's

just the midpoint. So the midpoint of 0 and 1 minus y is exactly 1 minus $y/2$. So that's a nice way of verifying that this answer is actually correct.

Now, the second part of part D asks us to do a little bit more. We have to use the total expectation theorem in order to somehow write the expectation of x in terms of the expectation of y . So the first thing we'll do is use the total expectation theorem.

So the total expectation theorem is just saying, well, we can take these conditional expectations. And now, we can integrate this by the marginal density of y , then we'll get the actual expectation of x . You can think of it as just kind of applying the law of iterated expectations as well. So this integral is going to look like this.

You take the conditional expectation. So this is the expectation of x if y were equal to little y . And now, what is that probability? Well, now we just multiply that by the density of y at that actual value of little y .

And we integrate with respect to y . Now, we've already calculated what this conditional expectation is. It's $1 - y/2$. So let's plug that in. $1 - y/2$ times the marginal of y .

There's a couple ways of attacking this problem now. One way is, we can actually just plug in that marginal of y . We've already calculated that out in part B. And then we can do this integral and calculate out the expectation.

But maybe we don't really want to do so much calculus. So let's do what the problem says and try a different approach. So what the problem suggests is to write this in terms of the expectation of y . And what is the expectation of y ?

Well, the expectation of y is going to look something like the integral of y times the marginal of y . So let's see if we can identify something like that and pull it out. Well, yeah, we actually do have that. We have y times the marginal of y , integrated.

So let's isolate that. So besides that, we also have this. We have the integral of the first term, is $1/2$ times the marginal of y . And then the second term is minus $1/2$ times the integral of y of dy .

This is just me splitting this integral up into two separate integrals. Now, we know what this is. The $1/2$ we can pull out. And then the rest of it is just the integral of a marginal of a density from minus infinity to infinity. And by definition, that has to be equal to 1. So this just gives us a $1/2$.

And now, what is this? We get a minus $1/2$. And now this, we already said that is the expectation of y . So what we have is the expectation of y .

So in the second part of this part D, we've expressed the expectation of x in terms of the expectation of y . Now, maybe that seems like that's not too helpful, because we don't know what either of those two are. But if we think about this problem, and as part E suggests, we can see that there's symmetry in this problem, because x and y are essentially symmetric.

So imagine this is x equals y . There's symmetry in this problem, because if you were to swap the roles of x and y , you would have exactly the same joint PDF. So what that suggests is that by symmetry then, it must be that the expectation of x and the expectation of y are exactly the same.

And that is using the symmetry argument. And that helps us now, because we can plug that in and solve for expectation of x . So expectation of x is $1/2$ minus $1/2$ expectation of x . So we have $3/2$ expectation of x equals $1/2$.

So expectation of x equals $1/3$. And of course, expectation of y is also $1/3$. And so it turns out that the expectation is around there. So this problem had several parts.

And it allowed us to start out from just a raw joint distribution, calculate marginals, calculate conditionals, and then from there, calculate all kinds of conditional expectations and expectations. And a couple of important points to remember are, when you do these joint distributions, it's very important to consider where values are valid.

So you have to keep in mind when you write out these conditional PDFs and joint PDFs and marginal PDFs, what ranges the formulas you calculated are valid for. And that also translates to when you're calculating expectations and such. When you have integrals, you need to be very careful about the limits of your integration, to make sure that they line up with the range where the values are actually valid.

And the last thing, which is kind of unrelated, but it is actually a common tool that's used in a lot of problems is, when you see symmetry in these problems, that can help a lot, because it will simplify things and allow you to use facts like these to help you calculate what the final answer is. Of course, this is also comes along with practice.

You may not immediately see that there could be a symmetry argument that will help with this problem. But with practice, when you do more of these problems, you'll eventually build up that kind of--

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

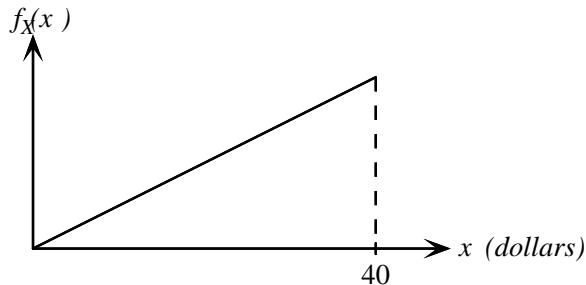
Problem Set 5
Due October 18, 2010

1. Random variables X and Y are distributed according to the joint PDF

$$f_{X,Y}(x,y) = \begin{cases} ax, & \text{if } 1 \leq x \leq y \leq 2, \\ 0, & \text{otherwise.} \end{cases}$$

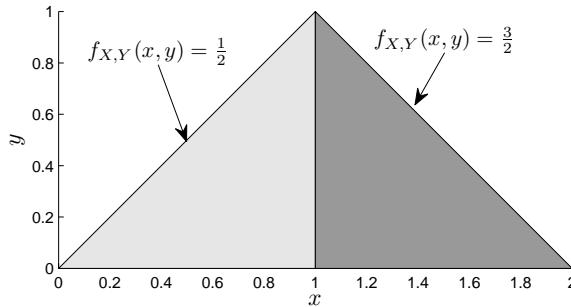
- (a) Evaluate the constant a .
- (b) Determine the marginal PDF $f_Y(y)$.
- (c) Determine the expected value of $\frac{1}{X}$, given that $Y = \frac{3}{2}$.

2. Paul is vacationing in Monte Carlo. The amount X (in dollars) he takes to the casino each evening is a random variable with the PDF shown in the figure. At the end of each night, the amount Y that he has on leaving the casino is uniformly distributed between zero and twice the amount he took in.



- (a) Determine the joint PDF $f_{X,Y}(x,y)$. Be sure to indicate what the sample space is.
- (b) What is the probability that on any given night Paul makes a positive profit at the casino? Justify your reasoning.
- (c) Find and sketch the probability density function of Paul's profit on any particular night, $Z = Y - X$. What is $\mathbf{E}[Z]$? Please label all axes on your sketch.

3. X and Y are continuous random variables. X takes on values between 0 and 2 while Y takes on values between 0 and 1. Their joint pdf is indicated below.



- (a) Are X and Y independent? Present a convincing argument for your answer.
 (b) Prepare neat, fully labelled plots for $f_X(x)$, $f_{Y|X}(y | 0.5)$, and $f_{X|Y}(x | 0.5)$.
 (c) Let $R = XY$ and let A be the event $X < 0.5$. Evaluate $\mathbf{E}[R | A]$.
 (d) Let $W = Y - X$ and determine the cumulative distribution function (CDF) of W .
4. **Signal Classification:** Consider the communication of binary-valued messages over some transmission medium. Specifically, any message transmitted between locations is one of two possible symbols, 0 or 1. Each symbol occurs with equal probability. It is also known that any numerical value sent over this wire is subject to distortion; namely, if the value X is transmitted, the value Y received at the other end is described by $Y = X + N$ where the random variable N represents additive noise that is independent of X . The noise N is normally distributed with mean $\mu = 0$ and variance $\sigma^2 = 4$.

- (a) Suppose the transmitter encodes the symbol 0 with the value $X = -2$ and the symbol 1 with the value $X = 2$. At the other end, the received message is decoded according to the following rules:
- If $Y \geq 0$, then conclude the symbol 1 was sent.
 - If $Y < 0$, then conclude the symbol 0 was sent.

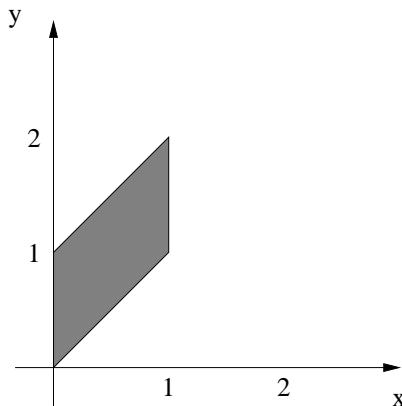
Determine the probability of error for this encoding/decoding scheme. Reduce your calculations to a single numerical value.

- (b) In an effort to reduce the probability of error, the following modifications are made. The transmitter encodes the symbols with a repeated scheme. The symbol 0 is encoded with the vector $\bar{X} = [-2, -2, -2]^\top$ and the symbol 1 is encoded with the vector $\bar{X} = [2, 2, 2]^\top$. The vector $\bar{Y} = [Y_1, Y_2, Y_3]^\top$ received at the other end is described by $\bar{Y} = \bar{X} + \bar{N}$. The vector $\bar{N} = [N_1, N_2, N_3]^\top$ represents the noise vector where each N_i is a random variable assumed to be normally distributed with mean $\mu = 0$ and variance $\sigma^2 = 4$. Assume each N_i is independent of each other and independent of the X_i 's. Each component value of \bar{Y} is decoded with the same rule as in part (a). The receiver then uses a majority rule to determine which symbol was sent. The receiver's decoding rules are:

- If 2 or more components of \bar{Y} are greater than 0, then conclude the symbol 1 was sent.
- If 2 or more components of \bar{Y} are less than 0, then conclude the symbol 0 was sent.

Determine the probability of error for this modified encoding/decoding scheme. Reduce your calculations to a single numerical value.

5. The random variables X and Y are described by a joint PDF which is constant within the unit area quadrilateral with vertices $(0, 0)$, $(0, 1)$, $(1, 2)$, and $(1, 1)$.



- (a) Are X and Y independent?
 (b) Find the marginal PDFs of X and Y .
 (c) Find the expected value of $X + Y$.
 (d) Find the variance of $X + Y$.
6. A defective coin minting machine produces coins whose probability of heads is a random variable P with PDF

$$f_P(p) = \begin{cases} 1 + \sin(2\pi p), & \text{if } p \in [0, 1], \\ 0, & \text{otherwise.} \end{cases}$$

In essence, a specific coin produced by this machine will have a fixed probability $P = p$ of giving heads, but you do not know initially what that probability is. A coin produced by this machine is selected and tossed repeatedly, with successive tosses assumed independent.

- (a) Find the probability that the first coin toss results in heads.
 (b) Given that the first coin toss resulted in heads, find the conditional PDF of P .
 (c) Given that the first coin toss resulted in heads, find the conditional probability of heads on the second toss.

G1[†]. Let C be the circle $\{(x, y) \mid x^2 + y^2 \leq 1\}$. A point a is chosen randomly on the boundary of C and another point b is chosen randomly from the interior of C (these points are chosen independently and uniformly over their domains). Let R be the rectangle with sides parallel to the x - and y -axes with diagonal ab . What is the probability that no point of R lies outside of C ?

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Problem Set 5: Solutions

1. (a) Because of the required normalization property of any joint PDF,

$$1 = \int_{x=1}^2 \left(\int_{y=x}^2 ax \, dy \right) dx = \int_{x=1}^2 ax(2-x) \, dx = a \left(2^2 - 1^2 - \frac{2^3}{3} + \frac{1^3}{3} \right) = \frac{2}{3}a$$

so $a = 3/2$.

- (b) For $1 \leq y \leq 2$,

$$f_Y(y) = \int_1^y ax \, dx = \frac{a}{2}(y^2 - 1) = \frac{3}{4}(y^2 - 1),$$

and $f_Y(y) = 0$ otherwise.

- (c) First notice that for $1 \leq x \leq 3/2$,

$$f_{X|Y}(x | 3/2) = \frac{f_{X,Y}(x, 3/2)}{f_Y(3/2)} = \frac{(3/2)x}{\frac{3}{4} \left(\left(\frac{3}{2}\right)^2 - 1^2 \right)} = \frac{8x}{5}.$$

Therefore,

$$\mathbf{E}[1/X | Y = 3/2] = \int_1^{3/2} \frac{1}{x} \frac{8x}{5} \, dx = 4/5.$$

2. (a) By definition $f_{X,Y}(x, y) = f_X(x)f_{Y|X}(y | x)$. $f_X(x) = ax$ as shown in the graph. We have that

$$1 = \int_0^{40} ax \, dx = 800a.$$

So $f_X(x) = x/800$. From the problem statement $f_{Y|X}(y | x) = \frac{1}{2x}$ for $y \in [0, 2x]$. Therefore,

$$f_{X,Y}(x, y) = \begin{cases} 1/1600, & \text{if } 0 \leq x \leq 4 \text{ and } 0 < y < 2x, \\ 0, & \text{otherwise.} \end{cases}$$

- (b) Paul makes a positive profit if $Y > X$. This occurs with probability

$$\mathbf{P}(Y > X) = \int \int_{y>x} f_{X,Y}(x, y) \, dy \, dx = \int_0^{40} \int_x^{2x} \frac{1}{1600} \, dy \, dx = \frac{1}{2}.$$

We could have also arrived at this answer by realizing that for each possible value of X , there is a $1/2$ probability that $Y > X$.

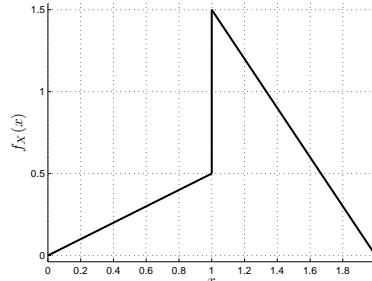
- (c) The joint density function satisfies $f_{X,Z}(x, z) = f_X(x)f_{Z|X}(z|x)$. Since Z is conditionally uniformly distributed given X , $f_{Z|X}(z | x) = \frac{1}{2x}$ for $-x \leq z \leq x$. Therefore, $f_{X,Z}(x, z) = 1/1600$ for $0 \leq x \leq 40$ and $-x \leq z \leq x$. The marginal density $f_z(z)$ is calculated as

$$f_Z(z) = \int_x^{40} f_{X,Z}(x, z) \, dx = \int_{x=|z|}^{40} \frac{1}{1600} \, dx = \begin{cases} \frac{40-|z|}{1600}, & \text{if } |z| < 40, \\ 0, & \text{otherwise.} \end{cases}$$

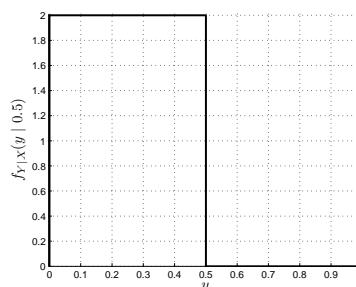
3. (a) In order for X and Y to be independent, any observation of X should not give any information on Y . If X is observed to be equal to 0, then Y must be 0.

In other words, $f_{Y|X=0}(y | 0) \neq f_Y(y)$. Therefore, X and Y are not independent.

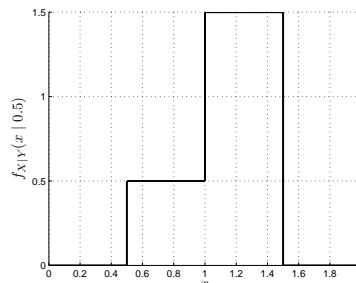
$$(b) f_X(x) = \begin{cases} x/2, & \text{if } 0 \leq x \leq 1, \\ -3x/2 + 3, & \text{if } 1 < x \leq 2, \\ 0, & \text{otherwise.} \end{cases}$$



$$f_{Y|X}(y | 0.5) = \begin{cases} 2, & \text{if } 0 \leq y \leq 1/2, \\ 0, & \text{otherwise.} \end{cases}$$



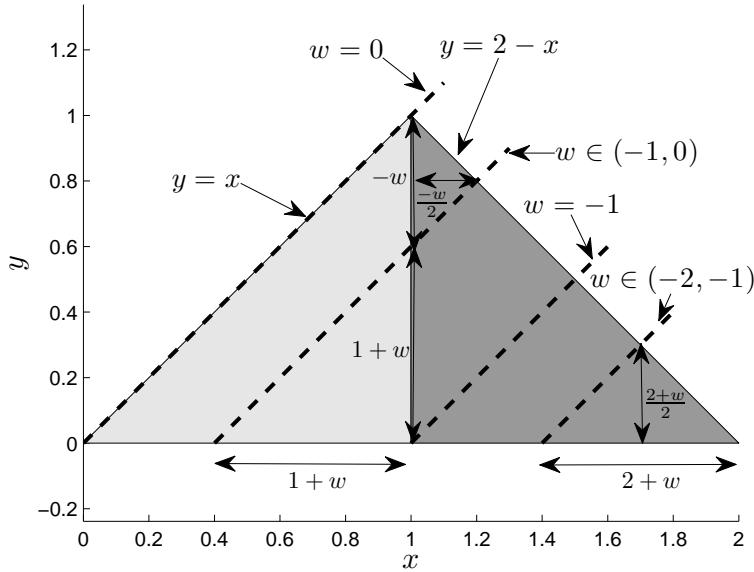
$$f_{X|Y}(x | 0.5) = \begin{cases} 1/2, & \text{if } 1/2 \leq x \leq 1, \\ 3/2, & \text{if } 1 < x \leq 3/2, \\ 0, & \text{otherwise.} \end{cases}$$



- (c) The event A leaves us with a right triangle with a constant height. The conditional PDF is then $1/\text{area} = 8$. The conditional expectation yields:

$$\begin{aligned} \mathbf{E}[R | A] &= \mathbf{E}[XY | A] \\ &= \int_0^{0.5} \int_y^{0.5} 8xy \, dx \, dy \\ &= 1/16. \end{aligned}$$

- (d) The CDF of W is $F_W(w) = \mathbf{P}(W \leq w) = \mathbf{P}(Y - X \leq w) = \mathbf{P}(Y \leq X + w)$. $\mathbf{P}(Y \leq X + w)$ can be computed by integrating the area below the line $Y = X + w$ for all possible values of w . The lines $Y = X + w$ are shown below for $w = 0$, $w = -1/2$, $w = -1$ and $w = -3/2$. The probabilities of interest can be calculated by taking advantage of the uniform PDF over the two triangles. Remember to multiply the areas by the appropriate joint density $f_{X,Y}(x,y)$! Take note that there are 4 regions of interest: $w < -2$, $-2 \leq w \leq -1$, $-1 < w \leq 0$ and $w > 0$.



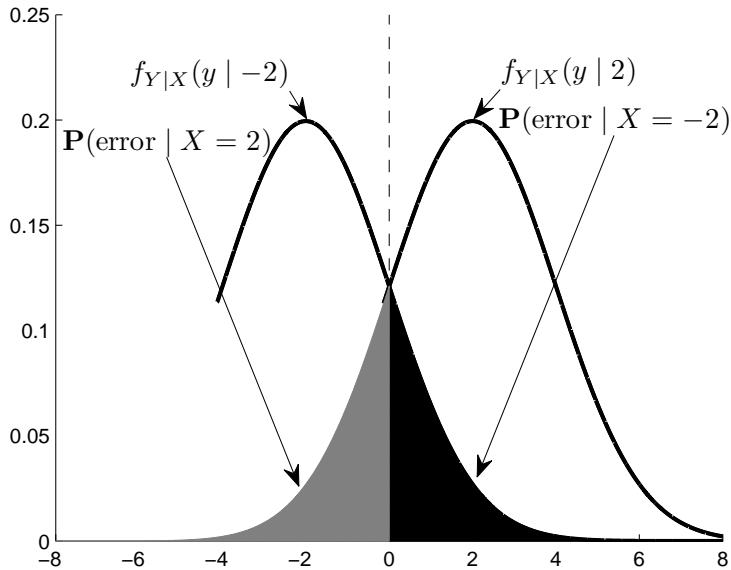
The CDF of W is

$$F_W(w) = \begin{cases} 0, & \text{if } w < -2, \\ 3/2 \cdot 1/2(2+w)^2/2, & \text{if } -2 \leq w \leq -1, \\ 1/2 \cdot 1/2(1+w)^2 + 3/2 \cdot (1/2 \cdot 1 \cdot 1 - 1/2(-w/2 \cdot -w)), & \text{if } -1 < w \leq 0, \\ 1, & \text{if } w > 0 \end{cases}$$

$$= \begin{cases} 0, & \text{if } w < -2, \\ 3/8 \cdot (2+w)^2, & \text{if } -2 \leq w \leq -1, \\ 1/8 \cdot (-w^2 + 4w + 8), & \text{if } -1 < w \leq 0, \\ 1, & \text{if } w > 0. \end{cases}$$

As a sanity check, $F_W(-\infty) = 0$ and $F_W(+\infty) = 1$. Also, $F_W(w)$ is continuous at $w = -2$ and at $w = -1$.

4. (a) If the transmitter sends the 0 symbol, the received signal is a normal random variable with a mean of -2 and a variance of 4 . In other words, $f_{Y|X}(y | -2) = \mathcal{N}(-2, 4)$.
 Also, $f_{Y|X}(y | 2) = \mathcal{N}(2, 4)$ These conditional pdfs are shown in the graph below.



The probability of error can be found using the total probability theorem.

$$\begin{aligned}
 \mathbf{P}(\text{error}) &= \mathbf{P}(\text{error} | X = -2)\mathbf{P}(X = -2) + \mathbf{P}(\text{error} | X = 2)\mathbf{P}(X = 2) \\
 &= \frac{1}{2}(\mathbf{P}(Y \geq 0 | X = -2) + \mathbf{P}(Y < 0 | X = 2)) \\
 &= \frac{1}{2}(\mathbf{P}(N \geq 2 | X = -2) + \mathbf{P}(N < -2 | X = 2)) \\
 &= \frac{1}{2}(\mathbf{P}(N \geq 2) + \mathbf{P}(N < -2)) \\
 &= \frac{1}{2}(\mathbf{P}\left(\frac{N-0}{2} \geq \frac{2-0}{2}\right) + \mathbf{P}\left(\frac{N-0}{2} < \frac{-2-0}{2}\right)) \\
 &= \frac{1}{2}((1 - \Phi(1)) + (1 - \Phi(1))) \\
 &= 0.1587.
 \end{aligned}$$

- (b) With 3 components, the probability of error given an observation of X is the probability of decoding 2 or 3 of the components incorrectly. For each component, the probability of error is 0.1587. Therefore,

$$\begin{aligned}
 \mathbf{P}(\text{error} | \text{sent 0}) &= \binom{3}{2}(0.1587)^2(1 - 0.1587) + (0.1587)^3 \\
 &= 0.0676.
 \end{aligned}$$

By symmetry, $\mathbf{P}(\text{error} | \text{sent 1}) = \mathbf{P}(\text{error} | \text{sent 0})$.

Therefore, $\mathbf{P}(\text{error}) = \mathbf{P}(\text{error} | \text{sent 0})\mathbf{P}(\text{sent 0}) + \mathbf{P}(\text{error} | \text{sent 1})\mathbf{P}(\text{sent 1}) = 0.0676$.

5. (a) There are many ways to show that X and Y are not independent. One of the most intuitive arguments is that knowing the value of X limits the range of Y , and vice versa. For instance, if it is known in a particular trial that $X \geq 1/2$, the value of Y in that trial cannot be smaller

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
 (Fall 2010)

than $1/2$. Another way to prove that the two are not independent is to calculate the product of their expectations, and show that this is not equal to $\mathbf{E}[XY]$.

- (b) Applying the definition of a marginal PDF,

for $0 \leq x \leq 1$,

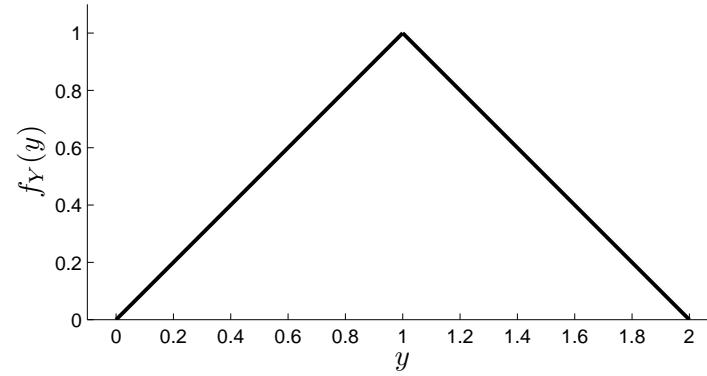
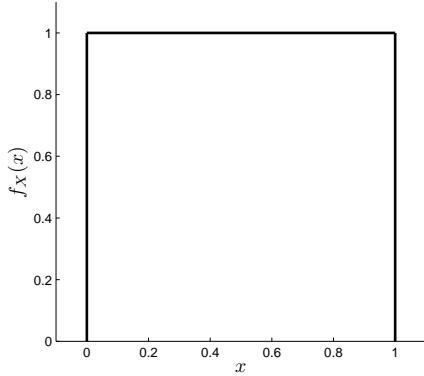
$$\begin{aligned} f_X(x) &= \int_y f_{X,Y}(x,y) dy \\ &= \int_x^{x+1} 1 dy \\ &= 1; \end{aligned}$$

for $0 \leq y \leq 1$,

$$\begin{aligned} f_Y(y) &= \int_x f_{X,Y}(x,y) dx \\ &= \int_0^y 1 dx \\ &= y; \end{aligned}$$

and for $1 \leq y \leq 2$,

$$\begin{aligned} f_Y(y) &= \int_x f_{X,Y}(x,y) dx \\ &= \int_{y-1}^1 1 dx \\ &= 2 - y. \end{aligned}$$



- (c) By linearity of expectation, the expected value of a sum is the sum of the expected values.
 By inspection, $\mathbf{E}[X] = 1/2$ and $\mathbf{E}[Y] = 1$.
 Thus, $\mathbf{E}[X + Y] = \mathbf{E}[X] + \mathbf{E}[Y] = 3/2$.

(d) The variance of $X + Y$ is

$$\mathbf{E}[(X + Y)^2] - \mathbf{E}[X + Y]^2 = \mathbf{E}[X^2] + 2\mathbf{E}[XY] + \mathbf{E}[Y^2] - (\mathbf{E}[X + Y])^2. \quad (1)$$

In part (c), $\mathbf{E}[X+Y]$ was computed, so only the other three expressions need to be calculated. First, the expected value of X^2 :

$$\mathbf{E}[X^2] = \int_0^1 x^2 \int_x^{x+1} 1 \, dy \, dx = \int_0^1 x^2 \, dx = 1/3.$$

Also, the expected value of Y^2 is

$$\mathbf{E}[Y^2] = \int_0^1 \int_x^{x+1} y^2 \, dy \, dx = \int_0^1 (3x^2 + 3x + 1)/3 \, dx = 7/6.$$

Finally, the expected value of XY is

$$\begin{aligned} \mathbf{E}[XY] &= \int_0^1 x \int_x^{x+1} y \, dy \, dx \\ &= \int_0^1 (2x^2 + x)/2 \, dx \, dy = 7/12. \end{aligned}$$

Substituting these into (1), we get $\text{var}(X + Y) = 1/3 + 7/6 + 7/6 - 9/4 = 5/12$.

Alternative (shortcut) solution to parts (c) and (d)

Given any value of X (in $([0,1])$), we observe that $Y - X$ takes values between 0 and 1, and is uniformly distributed. Since the conditional distribution of $Y - X$ is the same for every value of X in $[0,1]$, we see that $Y - X$ independent of X . Thus: (a) X is uniform, and (b) $Y = X + U$, where U is also uniform and independent of X . It follows that $\mathbf{E}[X + Y] = \mathbf{E}[2X + U] = 3/2$. Furthermore, $\text{var}(X + Y) = 4 \text{var}(X) + \text{var}(U) = 5/12$.

6. (a) Let A be the event that the first coin toss resulted in heads. To calculate the probability $\mathbf{P}(A)$, we use the continuous version of the total probability theorem:

$$\mathbf{P}(A) = \int_0^1 \mathbf{P}(A | P = p) f_P(p) \, dp = \int_0^1 p(1 + \sin(2\pi p)) \, dp,$$

which after some calculation yields

$$\mathbf{P}(A) = \frac{\pi - 1}{2\pi}.$$

- (b) Using Bayes rule,

$$\begin{aligned} f_{P|A}(p) &= \frac{\mathbf{P}(A | P = p) f_P(p)}{\mathbf{P}(A)} \\ &= \begin{cases} \frac{2\pi p(1 + \sin(2\pi p))}{\pi - 1}, & \text{if } 0 \leq p \leq 1, \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

(c) Let B be the event that the second toss resulted in heads. We have

$$\begin{aligned}\mathbf{P}(B | A) &= \int_0^1 \mathbf{P}(B | P = p, A) f_{P|A}(p) dp \\ &= \int_0^1 \mathbf{P}(B | P = p) f_{P|A}(p) dp \\ &= \frac{2\pi}{\pi - 1} \int_0^1 p^2(1 + \sin(2\pi p)) dp.\end{aligned}$$

After some calculation, this yields

$$\mathbf{P}(B | A) = \frac{2\pi}{\pi - 1} \cdot \frac{2\pi - 3}{6\pi} = \frac{2\pi - 3}{3\pi - 3} \approx 0.5110.$$

G1[†]. Let $a = (\cos \theta, \sin \theta)$ and $b = (b_x, b_y)$. We will show that no point of R lies outside C if and only if

$$|b| \leq |\sin \theta|, \quad \text{and} \quad |a| \leq |\cos \theta| \tag{2}$$

The other two vertices of R are $(\cos \theta, b_y)$ and $(b_x, \sin \theta)$. If $|b_x| \leq |\cos \theta|$ and $|b_y| \leq |\sin \theta|$, then each vertex (x, y) of R satisfies $x^2 + y^2 \leq \cos^2 \theta + \sin^2 \theta = 1$ and no points of R can lie outside of C . Conversely if no points of R lie outside C , then applying this to the two vertices other than a and b , we find

$$\cos^2 \theta + b^2 \leq 1, \quad \text{and} \quad a^2 + \sin^2 \theta \leq 1.$$

which is equivalent to 2.

These conditions imply that (b_x, b_y) lies inside or on C , so for any given θ , the probability that the random point $b = (b_x, b_y)$ satisfies (2) is

$$\frac{2|\cos \theta| \cdot 2|\sin \theta|}{\pi} = \frac{2}{\pi} |\sin(2\theta)|$$

and the overall probability is

$$\frac{1}{2\pi} \int_0^{2\pi} \frac{2}{\pi} |\sin(2\theta)| d\theta = \frac{4}{\pi^2} \int_0^{\pi/2} \sin(2\theta) d\theta = \frac{4}{\pi^2}$$

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

LECTURE 9

- **Readings:** Sections 3.4-3.5

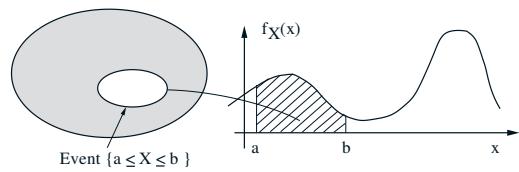
Outline

- PDF review
- Multiple random variables
 - conditioning
 - independence
- Examples

Summary of concepts

$p_X(x)$	$f_X(x)$
	$F_X(x)$
$\sum_x xp_X(x)$	$E[X] = \int x f_X(x) dx$
	$\text{var}(X)$
$p_{X,Y}(x, y)$	$f_{X,Y}(x, y)$
$p_{X A}(x)$	$f_{X A}(x)$
$p_{X Y}(x y)$	$f_{X Y}(x y)$

Continuous r.v.'s and pdf's



$$P(a \leq X \leq b) = \int_a^b f_X(x) dx$$

- $P(x \leq X \leq x + \delta) \approx f_X(x) \cdot \delta$
- $E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx$

Joint PDF $f_{X,Y}(x, y)$

$$P((X, Y) \in S) = \int \int_S f_{X,Y}(x, y) dx dy$$

- Interpretation:

$$P(x \leq X \leq x+\delta, y \leq Y \leq y+\delta) \approx f_{X,Y}(x, y) \cdot \delta^2$$

- Expectations:

$$E[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{X,Y}(x, y) dx dy$$

- From the joint to the marginal:

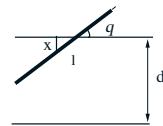
$$f_X(x) \cdot \delta \approx P(x \leq X \leq x + \delta) =$$

- X and Y are called **independent** if

$$f_{X,Y}(x, y) = f_X(x)f_Y(y), \quad \text{for all } x, y$$

Buffon's needle

- Parallel lines at distance d
Needle of length ℓ (assume $\ell < d$)
- Find $P(\text{needle intersects one of the lines})$



- $X \in [0, d/2]$: distance of needle midpoint to nearest line
- **Model:** X, Θ uniform, independent

$$f_{X,\Theta}(x, \theta) = \begin{cases} 1/d & 0 \leq x \leq d/2, 0 \leq \theta \leq \pi/2 \\ 0 & \text{otherwise} \end{cases}$$

- Intersect if $X \leq \frac{\ell}{2} \sin \Theta$

$$\begin{aligned} P\left(X \leq \frac{\ell}{2} \sin \Theta\right) &= \int \int_{x \leq \frac{\ell}{2} \sin \theta} f_X(x) f_\Theta(\theta) dx d\theta \\ &= \frac{4}{\pi d} \int_0^{\pi/2} \int_0^{(\ell/2) \sin \theta} dx d\theta \\ &= \frac{4}{\pi d} \int_0^{\pi/2} \frac{\ell}{2} \sin \theta d\theta = \frac{2\ell}{\pi d} \end{aligned}$$

Conditioning

- Recall

$$\mathbb{P}(x \leq X \leq x + \delta) \approx f_X(x) \cdot \delta$$

- By analogy, would like:

$$\mathbb{P}(x \leq X \leq x + \delta | Y \approx y) \approx f_{X|Y}(x | y) \cdot \delta$$

- This leads us to the **definition**:

$$f_{X|Y}(x | y) = \frac{f_{X,Y}(x, y)}{f_Y(y)} \quad \text{if } f_Y(y) > 0$$

- For given y , conditional PDF is a (normalized) "section" of the joint PDF

- If independent, $f_{X,Y} = f_X f_Y$, we obtain

$$f_{X|Y}(x | y) = f_X(x)$$

Joint, Marginal and Conditional Densities

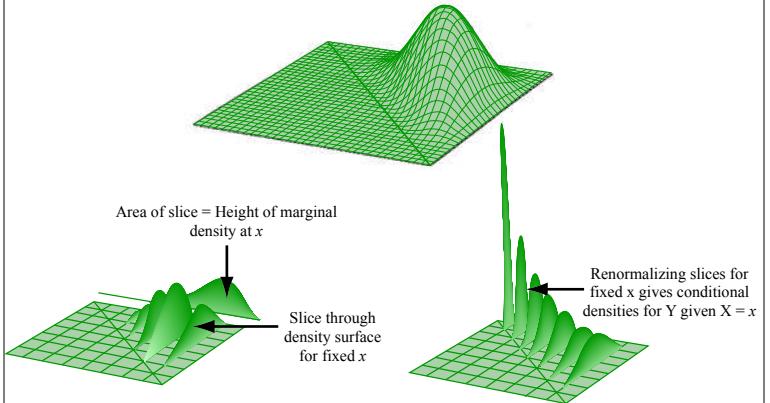
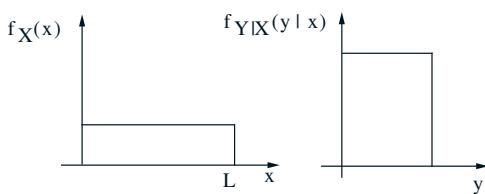


Image by MIT OpenCourseWare, adapted from *Probability*, by J. Pittman, 1999.

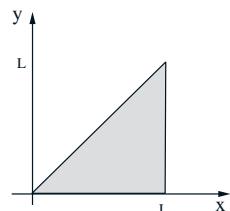
Stick-breaking example

- Break a stick of length ℓ twice:
break at X : uniform in $[0, 1]$;
break again at Y , uniform in $[0, X]$



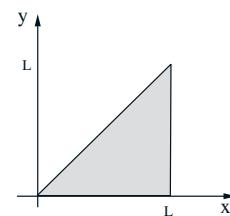
$$f_{X,Y}(x, y) = f_X(x)f_{Y|X}(y | x) =$$

on the set:



$$\mathbb{E}[Y | X = x] = \int y f_{Y|X}(y | X = x) dy =$$

$$f_{X,Y}(x, y) = \frac{1}{\ell x}, \quad 0 \leq y \leq x \leq \ell$$



$$\begin{aligned} f_Y(y) &= \int f_{X,Y}(x, y) dx \\ &= \int_y^\ell \frac{1}{\ell x} dx \\ &= \frac{1}{\ell} \log \frac{\ell}{y}, \quad 0 \leq y \leq \ell \end{aligned}$$

$$\mathbb{E}[Y] = \int_0^\ell y f_Y(y) dy = \int_0^\ell y \frac{1}{\ell} \log \frac{\ell}{y} dy = \frac{\ell}{4}$$

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Recitation 9
October 7, 2010

1. Let X be an exponential random variable with parameter $\lambda > 0$. Calculate the probability that X belongs to one of the intervals $[n, n + 1]$ with n odd.
2. (Example 3.13 of the text book, page 165) **Exponential Random Variable is Memoryless.** The time T until a new light bulb burns out is an exponential random variable with parameter λ . Ariadne turns the light on, leaves the room, and when she returns, t time units later, finds that the bulb is still on, which corresponds to the event $A = \{T > t\}$. Let X be the additional time until the bulb burns out. What is the conditional CDF of X , given the event A ?
3. Problem 3.23, page 191 in the text.
Let the random variables X and Y have a joint PDF which is uniform over the triangle with vertices $(0, 0)$, $(0, 1)$, and $(1, 0)$.
 - (a) Find the joint PDF of X and Y .
 - (b) Find the marginal PDF of Y .
 - (c) Find the conditional PDF of X given Y .
 - (d) Find $\mathbf{E}[X | Y = y]$, and use the total expectation theorem to find $\mathbf{E}[X]$ in terms of $\mathbf{E}[Y]$.
 - (e) Use the symmetry of the problem to find the value of $\mathbf{E}[X]$.
4. We have a stick of unit length, and we break it into three pieces. We choose randomly and independently two points on the stick using a uniform PDF, and we break the stick at these points. What is the probability that the three pieces we are left with can form a triangle?

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Recitation 9 Solutions
October 7, 2010

1. We first compute the probability that X is in interval $[n, n + 1]$ for an arbitrary nonnegative n . Then, we will add the probabilities for all odd positive integer values of n .

We could integrate the PDF of X over the given interval but we will use the CDF here. Using the CDF for the exponential random variable,

$$\begin{aligned}\mathbf{P}(n \leq X \leq n + 1) &= F_X(n + 1) - F_X(n) \\ &= \left(1 - e^{-\lambda(n+1)}\right) - \left(1 - e^{-\lambda n}\right) \\ &= e^{-\lambda n} \left(1 - e^{-\lambda}\right).\end{aligned}$$

Since the intervals are disjoint, we can sum this probability for all odd integers n to find the probability of interest:

$$\begin{aligned}\mathbf{P}(\{X \in [n, n + 1] \text{ for some odd } n\}) &= \sum_{n \text{ odd}} e^{-\lambda n} \left(1 - e^{-\lambda}\right) \\ &= \left(1 - e^{-\lambda}\right) \sum_{k=0}^{\infty} e^{-\lambda(2k+1)} \\ &= \left(1 - e^{-\lambda}\right) e^{-\lambda} \sum_{k=0}^{\infty} \left(e^{-2\lambda}\right)^k \\ &= \left(1 - e^{-\lambda}\right) e^{-\lambda} \frac{1}{1 - e^{-2\lambda}} \\ &= \left(1 - e^{-\lambda}\right) e^{-\lambda} \frac{1}{(1 - e^{-\lambda})(1 + e^{-\lambda})} \\ &= \frac{e^{-\lambda}}{1 + e^{-\lambda}}.\end{aligned}$$

2. See Example 3.13 in the textbook on page 165.
3. Problem 3.23, page 191 in text. See online solutions.
4. Problem 3.22, part (i), page 191 in text (see online solution).

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Tutorial 4
October 7/8, 2010

1. Let X and Y be Gaussian random variables, with $X \sim N(0, 1)$ and $Y \sim N(1, 4)$.

- Find $\mathbf{P}(X \leq 1.5)$ and $\mathbf{P}(X \leq -1)$.
- What is the distribution of $\frac{Y-1}{2}$?
- Find $\mathbf{P}(-1 \leq Y \leq 1)$.

2. Example 3.15, page 169 in text.

Ben throws a dart at a circular target of radius r . We assume that he always hits the target, and that all points of impact (x, y) are equally likely. Compute the joint PDF $f_{X,Y}(x, y)$ of the random variables X and Y and compute the conditional PDF $f_{X|Y}(x|y)$.

3. Problem 3.20, page 191 in text.

An absent-minded professor schedules two student appointments for the same time. The appointment durations are independent and exponentially distributed with mean thirty minutes. The first student arrives on time, but the second student arrives five minutes late. What is the expected value of the time between the arrival of the first student and the departure of the second student?

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2010

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Tutorial 4: Solutions

1. (a)

$$\begin{aligned}\mathbf{P}(X \leq 1.5) &= \Phi(1.5) \\ &= 0.9332.\end{aligned}$$

$$\begin{aligned}\mathbf{P}(X \leq -1) &= 1 - \mathbf{P}(X \leq 1) \\ &= 1 - \Phi(1) \\ &= 1 - 0.8413 \\ &= 0.1587.\end{aligned}$$

(b)

$$\begin{aligned}\mathbf{E}\left[\frac{Y-1}{2}\right] &= \frac{1}{2}(\mathbf{E}[Y] - 1) \\ &= 0.\end{aligned}$$

$$\begin{aligned}\text{var}\left(\frac{Y-1}{2}\right) &= \text{var}\left(\frac{Y}{2}\right) \\ &= \frac{1}{4}\text{var}Y \\ &= 1.\end{aligned}$$

Thus, the distribution of $\frac{Y-1}{2}$ is $\mathcal{N}(0, 1)$.

(c)

$$\begin{aligned}\mathbf{P}(-1 \leq Y \leq 1) &= \mathbf{P}\left(\frac{-1-1}{2} \leq \frac{Y-1}{2} \leq \frac{1-1}{2}\right) \\ &= \Phi(0) - \Phi(-1) \\ &= \Phi(0) - (1 - \Phi(1)) \\ &= 0.3413.\end{aligned}$$

2. Example 3.15, page 169 in text. See solutions in the text.

3. Problem 3.20, page 191 in text. See online solutions.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013 Transcript – Recitation: A Derived Distribution Example

Hi. In this problem we'll work through an example of calculating a distribution for a minute variable using the method of derived distributions. So in general, the process goes as follows. We know the distribution for some random variable X and what we want is the distribution for another random variable of Y , which is somehow related to X through some function g . So Y is a g of X .

And the steps that we follow-- we can actually just kind of summarize them using this four steps. The first step is to write out the CDF of Y . So Y is thing that we want. And what we'll do is we'll write out the CDF first. So remember the CDF is just capital F of y , y is the probability that random variable Y is less than or equal to some value, little y .

The next thing we'll do is, we'll use this relationship that we know, between Y and X . And we'll substitute in, instead of writing the random variable Y In here, we'll write it in terms of X . So we'll plug in for-- instead of Y , we'll plug-in X . And we'll use this function g in order to do that.

So what we have now is that up to here, we would have that the CDF of Y is now the probability that the random variable X is less than or equal to some value, little y . Next what we'll do is we'll actually rewrite this probability as a CDF of X . So the CDF of X , remember, would be-- F of x is that the probability of X is less than or equal to some little x .

And then once we have that, if we differentiate this-- when we differentiate the CDF of X , we get the PDF of X . And what we presume is that we know this PDF already. And from that, what we get is, when we differentiate this thing, we get the PDF of Y . So through this whole process what we get is, we'll get the relationship between the PDF of Y and the PDF of X . So that is the process for calculating the PDF of Y using X .

So let's go into our specific example. In this case, what we're told is that X , the one that we know, is a standard normal random variable. Meaning that it's mean 0 and variance 1. And so we know the form of the PDF. The PDF of x is this, 1 over square root of $2\pi e$ to the minus x squared over 2 .

And then the next thing that we're told is this relationship between X and Y . So what we're told is, if X is negative, then Y is minus X . If X is positive, then Y is the square root of X . So this is a graphical its representation of the relationship between X and Y .

All right, so we have everything that we need. And now let's just go through this process and calculate what the PDF of Y is. So the first thing we do is we write out the PDF of Y . So the PDF of Y is what we've written. It's the probability that the random variable Y is less than or equal to some little y .

Now the next step that we do is we have to substitute in, instead of in terms of Y, we want to substitute it in terms of X. Because we actually know stuff about X, but we don't know anything about Y. So what is the probability that Y, the random variable Y, is less than or equal to some little y?

Well, let's go back to this relationship and see if we can figure that out. So let's pretend that here is our little y. Well, if the random variable Y is less than or equal to little y, it has to be underneath this horizontal line. And in order for it to be underneath this horizontal line, that means that X has to be between this range. And what is this range? This range goes from minus Y to Y squared.

So why is that? It's because in this portion X and Y are related as, Y is negative X and here it's Y is square root of X. So if X is Y squared, then Y would be Y. If X is negative Y, then Y would be Y. All right, so this is the range that we're looking for.

So if Y, the random variable Y is less than or equal to little y, then this is the same as if the random variable X is between negative Y and Y squared. So let's plug that in. This is the same as the probability that X is between negative Y and Y squared.

So those are the first two steps. Now the third step is, we have to rewrite this as the CDF of x. So right now we have it in terms of a probability of some event related to X. Let's actually transform that to be explicitly in terms of the CDF of X. So how do we do that?

Well, this is just the probability that X is within some range. So we can turn that into the CDF by writing it as a difference of two CDFs. So this is the same as the probability that X is less than or equal to Y squared minus the probability that X is less than or equal to negative Y.

So in order to find the probability that X is between this range, we take the probability that it's less than Y squared, which is everything here. And then we subtract that probability that it's less than Y, negative Y. So what we're left with is just within this range.

So these actually are now exactly CDFs of X. So this is F of X evaluated at Y squared and this is F of X evaluated at negative Y. So now we've completed step three. And the last step that we need to do is differentiate.

So if we differentiate both sides of this equation with respect to Y, we'll get that the left side would get what we want, which is the PDF of Y. Now we differentiate the right side-- we'll have to invoke the chain rule. So the first thing that we do is, well, this is a CDF of X. So when we differentiate we'll get the PDF of X.

But then we also have invoke the chain rule for this argument inside. So the derivative of Y squared would give us an extra term, 2Y. And then similarly this would give us the PDF of X evaluated at negative Y plus the chain will give us an extra term of negative 1. So let's just clean this up a little bit.

So it's $2y F X$ squared plus $F X$ minus Y . All right, so now we're almost done. We've differentiated. We have the PDF of Y , which is what we're looking for. And we've written it in terms of the PDF of X . And fortunately we know what that is, so once we plug that in, then we're essentially done.

So what is the PDF? Well, the PDF of X evaluated at Y squared is going to give us 1 over square root of $2 \pi e$ to the minus-- so in this case, X is Y squared-- so we get Y to the fourth over 2 . And then we get another 1 over square root of $2 \pi e$ to the minus Y squared over 2 . OK, and now we're almost done.

The last thing that we need to take care of is, what is the range? Now remember, it's important when you calculate out PDFs to always think about the ranges where things are valid. So when we think about this, what is the range where this actually is valid? Well, Y , remember is related to X in this relationship. So as we look at this, we see that Y can never be negative.

Because no matter what X is, Y gets transformed into some non-negative version. So what we know is that this is now actually valid only for Y greater than 0 and for Y less than 0 , the PDF is 0 . So this gives us the final PDF of Y .

All right, so it seems like at first when you start doing these derived restriction problems that it's pretty difficult. But if we just remember that there are these pretty straightforward steps that we follow, and as long as you go through these steps and do them methodically, then you can actually come up with the solution for any of these problems. And one last thing to remember is to always think about what are the ranges where these things are valid? Because the relationship between these two random variables could be pretty complicated and you need to always be aware of when things are non-zero and when they are 0 .

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013

Transcript – Tutorial: Ambulance Travel Time

In this problem, we'll be looking at an ambulance that is traveling back and forth in interval of size 1. Say from 0 to 1.

At some point in time, there's an accident occurring, let's say at location x . And we'll assume the accident occurs in a random location so that x is uniformly distributed between 0 and 1. Now, at this point in time, let's say the ambulance turns out to be at location y . Again, we'll assume that y is a uniform random variable between 0 and 1, and also that x and y are independently distributed.

The question we're interested in answering is how long it would take an ambulance to respond to travel from point y to point x . Let's call this time T . And in particular, we want to know everything about distribution of T . For example, what is the CDF of T given by the probability of big T , less than or equal to little t , or the PDF, which is done by differentiating the CDF once we have it.

Now, to start, we'll express T you as a function of X and Y . Since we know that the ambulance travels at a speed V -- V meters or V units of distance per second-- then we can write that big T is simply equal to Y minus X , absolute value the distance between X and Y , divided by the speed at which the ambulance is traveling at, V . So now if we look at the probability of T less than or equal to little t , this is then equal to the probability that Y minus X divided by V less than or equal to little t .

We now take off the absolute value by writing the expression as negative vt less equal to Y minus X less equal to positive vt . Here we multiply v on the other side of t , and then took out the absolute value sign. As a final step, we'll also move X to the other side of inequalities by writing this as X minus vt less equal to y less equal to x plus vt .

To compute this quantity, we'll define a set A as a set of all points that satisfies this condition right here. In particular, it's a pair of all X and Y such that X minus vt less equal to little y less equal to X plus vt , and also that X is within 0 and 1, and so is Y .

So the set A will be the set of values we'll be integrating over. Now that we have A , we can express the above probability as the integral of all X and Y , this pair within the set A , integrating the PDF of f of X , Y , little x , little y .

Let's now evaluate this expression right here in a graphical way. On the right, we're plotting out what we just illustrated here, where the shaded region is precisely the set A . As we can see, this is a set of values of X and Y where Y is sandwiched between two lines, the upper one being X plus vt right here, and the lower line being X minus vt , right here. So these are the values that correspond to the set A .

Now that we have A, let's look f of x, y. We know that both x and y are uniform random variables between 0 and 1, and therefore, since they're independent, the probability density of x and y being at any point between 0 and 1 is precisely 1 over 1 squared, where 1 squared is the size of this square box right here.

So given this picture, all we need to do is to multiply by 1 over 1 squared the area of the region A. And depending on the value of T, we'll get different answers as right here. If T is less than 0, obviously, the area of A diminishes to nothing, so we get 0. If T is greater than 1 over V, the area of A fills up the entire square, and we get 1. Now, if T is somewhere in between 0 and 1 over v, we will have 1 over 1 squared, multiply by the area looking like something like that right here--the shaded region.

Now, if you wonder how we arrive at exactly this expression right here, here is a simple way to calculate it. What we want is 1 over 1 squared times the area A. Now, area A can be viewed as the entire square, 1 squared, minus whatever's not in area A, which is these two triangles right here. Now, each triangle has area 1/2, 1 minus vt squared. This multiply 2, and this, after some algebra, will give the answer right here.

At this point, we have obtained the probability of big T less equal to little t. Namely, we have gotten the CDF for T. And as a final step, we can also compute the probability density function for T. We'll call it little f of t. And we do so by simply differentiating the CDF in different regions of T.

To begin, we'll look at t between 0 and 1 over v right here at differentiating the expression right here with respect to t. And doing so will give us $2v$ over $1 - 2vt$ squared t over $1/v$ squared. And this applies to t greater or equal to 0, less than $1/v$. Now, any other region, either t less than 0 or t greater than $1/v$, we have a constant for the CDF, and hence its derivative will be 0. So this is for any other t . We call it otherwise.

Now, this completely characterized the PDF of big T, and hence, we've also finished a problem.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013

Transcript – Recitation: Inferring a Continuous Random Variable from a Discrete Measurement

In this problem, we're going to look at how to infer a continuous random variable from a discrete measurement. And the continuous random variable that we're interested in in this problem is q , which is given by this PDF. It's $6q$ times $1 - q$ for a q between 0 and 1 and 0 otherwise.

And here is a graph of what it looks like. It just kind of has this curve shape. And it's symmetric. And its peak is at $1/2$.

And the way to interpret q is q is the unknown bias of a coin. So the bias of a coin, the probability of heads of that coin is somewhere between 0 and 1. We're not sure exactly what it is.

And here is our, say, prior belief about how this random bias is distributed. And we're going try to infer what this bias is by flipping the coin and observing whether or not we got heads or tails. And because of that, the measurement, or the observation that we get, is discreet. Either we get heads, or we get tails. And we model that using a discrete random variable x , which is, in this case, it turns out it's just a Bernoulli random variable, either 0 or 1.

And the model that we have is that, if we knew what this bias q was, if we knew that it was a little q , then this coin, I mean, it behaves as if it was a coin with that bias. So the probability of getting heads, or the probability that x equals 1, is just equal to q , which is the bias. And the probability that it's equal to 0 is $1 - q$. So that's just our model of how the coin works.

We can also write this another way, as just more like a conditional PMF. So the conditional PMF of x , given q of little x , is q , if x is 1, $1 - q$, if x equals 0, and 0 otherwise. Just a more compact way of writing this.

All right, so what we want to do in this problem is find this conditional PDF. What is the conditional PDF of q given x ? So we observe what x is, either a 0 or 1. And we want to know now, given that information, given that measurement, what is the new distribution of q the bias of the coin?

And to do that, well, we apply Bayes' rule. And remember, Bayes' rule, it consists of several terms. The first one is the numerator, which is our prior initial belief, so which is just the regular PDF of q , times the conditional PMF of x given q . All right, so because we have a continuous variable that we want to infer from a discreet measurement, we use this variation of Bayes' rule, where we have a PDF here and a conditional PMF here. And the denominator is-- well, it's the PMF of x .

And of course, we can take this PMF of x , this denominator, and expand it, using the law of total probability where the PMF of x , you can think of it as you can get x with a combination of lots of different possible values of the bias q . And so we just calculate all those possibilities and

integrate. And what we want to integrate here is q , so we want to integrate-- remember to keep in mind the limits of integration. And this is just referenced by the limits of what the PDF of q is.

OK. All right, so now we're asked to find what this value is for x equals to 0 or 1 and for all values of q . And the values of q we care about are the ones between 0 and 1. So let's focus on the two different possibilities for x .

So the first one is, let's look at the case where x equals 1 first. And then now let's just plug-in what all these different terms should be. Well, the PDF of q we're given. And of course, we're looking here at q between 0 and 1, so within that range. The PDF of q is just $6q$ times 1 minus q .

And the conditional PMF of x where we know that from our model, because we're looking at the case where x equals 1. That conditional PMF is just q . And the denominator is really the same as the numerator, except you integrate it. So it's the integral from 0 to 1 of $6q$ times 1 minus q times q dq .

OK. And now we can simplify this. So under the numerator, we have integral-- sorry, $6q$ squared times 1 minus q . And then the bottom we have the integral of $6q$ squared minus q cubed, d cubed from 0 to 1. And now this is just some calculus now. So we still have the numerator $6q$ squared times 1 minus q . The denominator, we have $2q$ cubed-- that would give us the $6q$ squared term-- minus $6/4 q$ to the fourth. And we integrate that from 0 to 1.

OK. And what does that give us? We get $6q$ squared 1 minus q still on the top. And the bottom, we get-- well the 0-- the case where it's 0, it's just 0. The case where it's 1, it's 2 minus $3/2$, so it's $1/2$. So really, it just becomes $12 q$ squared 1 minus q . And of course, this only true when q is between 0 and 1.

All right, so the case where it's equal to 1, we have our answer. And it turns out that, if you plot this, what does it look like? It looks something like this where the peak is now at $2/3$.

So how do you interpret this? The interpretation is that what you have is you observe that you've got, actually, heads on this toss. So that is evidence that the bias of the coin is relatively high.

So it's relatively more likely to get heads with this coin. So q , in that case, you would believe that it's more likely to be higher than $1/2$. And it turns out that, when you go through this reasoning and the Bayes' rule, what you get is that it looks like this. And the peak is now at $2/3$.

And you can repeat this exercise now with the case where x is 0. So when x is 0, you still get the same term here, $6q$ 1 minus q , but the conditional PMF is now the-- you want the conditional PMF when x equals 0, which is now 1 minus q . So you get 1 minus q here.

And now this term becomes $6q$ times 1 minus q squared. And so really, the bottom is also $6q$ times 1 minus q squared dq . And if you go through the same sort of calculus, what you get is that, in this case, the answer is $12q$ 1 minus q squared.

So let me rewrite what the first case was. The first case, when x equals 1 was equal to $12q$ squared 1 minus q . So they look quite similar, except that this one has q squared, this one has 1 minus q squared. And if you take this one, the case where you observe a 0, and you plot that, it turns out it looks something like this. And this actually doesn't look like it, but it should be the peak is at $1/3$.

And so you notice, first of all, that these are symmetric. There's some symmetry going on. And this interpretation in this case is that, because you observed 0, which corresponds to observing tails, that gives you evidence that the bias of the coin is relatively low, or the probability of getting heads with this coin is relatively low, which pushes your belief about q towards the smaller values.

OK, so it turns out that this distribution, this distribution, and the original distribution of q , all fall under family of distributions called the beta distribution. And they're parameterized by a couple of parameters. And it's used frequently to model things like the bias of a coin, or anything that's a random variable that's bounded between 0 and 1.

OK, so this problem allowed us to get some more exercise with Bayes' rule, this continuous discrete version of Bayes' rule. And if you go through all the steps, you'll find that it's relatively straightforward to go through the formula and plug in the different parts of the terms and go through a little bit of calculus and find the answer. And it's always good to go back, once you find the answer, to look at it a little bit and make sure that actually makes sense. So you can convince yourself that it's not something that looks--

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013

Transcript – Recitation: Inferring a Discrete Random Variable from a Continuous Measurement

Hi. In this problem, we're going to look at how to infer a discrete random variable from a continuous measurement. And really, what it's going to give us is some practice working with a variation of Bayes' rule. So the problem tells us that we have a discrete random variable x with this PMF. It is 1 with probability P , minus 1 with probability $1 - P$, and 0 otherwise. So here is just a diagram of this PMF.

And then we also have another random variable, y , which is continuous. And its PDF is given by this. It's $\frac{1}{2} \lambda e^{-\lambda|y|}$. And so this may look familiar. It looks kind of like an exponential. And in fact, it's just a two-sided exponential. That's flattened by a factor of $\frac{1}{2}$. And this is what it looks like, kind of like a tent that goes on both ways.

And then we have a random variable z , which is equal to the sum of x and y . And the problem is going to be figuring out what x is, based on an observed value of what z is. So because x is discrete and y is random-- sorry, x is discrete, and y is continuous, z is also going to be continuous. So our measurement is z , which is continuous. And we want to infer x , which is discrete.

So the problem asked us to find is this what is the probability that x equals 1, given that z is a little z . And you can write this another way, just as a conditional PMF as well. It's the conditional PMF of x , given z , evaluated to 1 conditioned on little z .

All right, so now let's apply the correct variation of Bayes' rule. So remember, it's going to be this, the probability that x equals 1, or the PMF of x evaluated to 1, times the-- you flip this conditioning. So now it's going to be a conditional PDF of z , since z is continuous. It's going to be a conditional PDF of z , given x , evaluated at some little z condition on x being 1.

And the bottom is the conditional PDF-- or sorry, just the regular PDF of z . And of course, we can rewrite this denominator. Remember, the denominator is always just-- you can use the law of total probability and rewrite it. And one of the terms is going to be exactly the same as the numerator.

So one of the ways that z can be some little z is it's in combination with x being equal to 1. And the probability of that is exactly the same thing as the numerator. And the other way is if x is equal to negative 1. And that gives us this second term.

All right. And now let's just fill in what all these different terms are. So with the PMF of x evaluated at 1, that is just P . What is the conditional PDF of z , given that x is equal to 1? Well, that takes a little bit more work.

Given that x is 1, then z is just going to be-- so if x equals 1, then z is just y plus 1, which means that you can just imagine taking y -- this is what y is, the distribution of y -- and just adding 1 to it, which, in this diagram, would amount to shifting it over by one. So now, it would look like this, the distribution. And algebraically, all you would do is just change this thing in the absolute value to y minus 1. That amounts to shifting it over to the right by one.

All right. So what is that? That's just $1/2 \lambda e^{-\lambda y}$ sorry, not y , z -- z minus 1. And the denominator, well, the first term is going to be exactly the same. It's just also $P 1/2 \lambda e^{-\lambda z}$.

What about the second term? The second term, first we need to figure out what is the PMF of x evaluated at a negative 1. Or in other words, what's the probability that x is negative 1? That is given to us by the PMF. It's $1 - P$.

And then the second part is, what is the conditional PDF of z , given that x is negative 1? Well, we can just do the same sort of trick here. If x is negative 1, then z is just y minus 1.

In which case, the PDF of z would just look like this. You're shifted to the left by one now. And now what you have to do is change this into a plus 1. So this conditional PDF would be $1/2 \lambda e^{-\lambda |z+1|}$.

All right, so this looks pretty messy. And we can try to simplify things a little bit. So we can get rid of these $1/2 \lambda$ s. And then we can multiply the numerator and the denominator by the same term. Let's multiply it by $e^{\lambda z}$.

So what we're going to do is try to cancel out some of these exponential terms. So that will cancel out this top term. So all we have in the numerator now is just P . It will also cancel out this exponential in the denominator. And then we'll have to change this here, because it'll have an extra $e^{\lambda z}$.

All right, now let's rewrite this. And what we get is $P e^{-\lambda |z+1|}$. OK, so that is pretty much as far as you can go in terms of simplifying it.

And now the question is, are we comfortable with this answer? And it helps always to try to interpret it a little bit, to make sure that it makes intuitive sense. And one way to do that is to try to-- some of the limiting cases of what some of the parameters can be.

So in this case, the parameters are P and λ . So P is the parameter related to x . And λ is the parameter related to y . So let's try to see if it makes sense under some limiting cases.

The first one we want to think about is when P goes to 0. So if P goes to 0, what happens to our answer? Well, the numerator is 0, this is 0, this is 1. But it doesn't matter, because the numerator is 0. So in this case, this would go to 0.

Now does that make sense? Well, what does that mean when P goes to 0? When P goes to 0, that means that the probability that x is equal to 1 is 0. So even without thinking about y or z , there is already a 0 probability that x is equal to 1.

Now this whole calculation, what we found is, well, if I had some more information, like what z is, does that help me find out what the probability of x being 1 is? Well, no matter what z tells me, I know for a fact that x can't be 1, because P is 0. So this posterior, or this conditional probability, should also be 0, because there's just no way that x can be 1. So in this case, this formula does check out.

Now let's think about another case where P goes to 1. If P goes to 1, that means that X is for sure going to be 1. And it can't be anything else. In which case, what does our formula tell us?

Well, this numerator is 1. This term is 1. 1 minus 1 is 0. So the second term gets zeroed out, and the answer is just $1/1$ is 1.

So what does this tell us? This tells us that if I know beforehand that x is for sure equal to 1, then, if I now give myself more information and condition on what I observe for z , that shouldn't change anything else. I should still know for sure that x is equal to 1. So the probability of this conditional probability should still be equal to 1. And it does, so our formula also works in this case.

Now let's think about lambda. What about when lambda goes to 0? Well, when lambda goes to 0, that's a little harder to visualize. But really, what would happen is that you can imagine this distribution getting shallower, shallower and shallower, lower and lower, so that it's like it is kind of flat and goes on forever.

And so what this tells you is that, basically, y -- this is the distribution y -- so when lambda goes to 0, that tells you that y has a really flat and kind of short distribution. And so what does our formula tell us in this case? Well, when lambda goes to 0, this exponent is equal to 0.

And so e to the 0 is 1. So we get P over P plus 1 minus P , which is just 1. So the answer here, our formula will give us an answer of P.

So what does that tell us? That tells us that, in this case, if lambda goes to 0, then our posterior probability, the probability that x equals 1 conditioned on z being some value, conditioned on our continuous measurement, is still P . So the prior, or the original probability for x being equal to 1 is P . And with this additional continuous measurement, our guess of the probability that x equal to 1 is still P . So it hasn't changed.

So basically, it's telling us that this additional information was not informative. It didn't actually help us change our beliefs. And so why is that?

Well, one way to think about it is that, because the distribution of y looks like this, is very flat and it could be anything, then, if you observe some value of z , then it could be that that was due to the fact that it was x equal to 1 plus some value of y that made z equal to that value. Or it

could have just as equally been likely that x equal to negative 1 y equals to some other value that made it equal to z .

And so, essentially, it's z -- because y has a shape, it can be likely to take on any value that complements either x being equal to 1 or x equal being to negative 1, to make z equal to whatever the value it is that you observe. And so because of that, in this case, y is not very informative. And so this probability is still just equal to P .

Now the last case is when lambda goes to infinity. And now we have to break it down into the two other cases now. The first case is when-- lets write this over here-- when lambda goes to infinity.

The first case, it depends on what this value is, the sine of this value. If this value, the absolute value of z plus 1 minus the absolute value of z minus 1, if that's positive, then, because lambda goes to infinity and you have a negative sign, then this entire exponential term will go to 0. In which case, the second term goes to 0. And the answer is P/P , or is 1. And so if absolute value of z plus 1 minus absolute value of z minus 1 is greater than 0, then the answer is 1.

But in the other case, if this term in the exponent, if it's actually negative, if it's negative, then this negative sign turns to a positive, and lambda goes to infinity. And so this term blows up, and it dominates everything else. And so the denominator goes to infinity. The numerator is fixed at P , so this entire expression would go to 0.

OK, so now let's try to interpret this case. Let's start with the first one. When is it that absolute value of z plus 1 minus absolute value of z minus 1 is greater than 0? Or you can also rewrite this as absolute value of z plus 1 is greater than absolute value of z minus 1. Well, when is that case?

Well, it turns out, if you think about it, this is only true if z is positive. If z is positive, then adding 1-- let me draw a line here, and if this is 0-- if z is positive, something here, adding 1 to it and taking the absolute value-- the absolute value doesn't do anything-- but you will get something bigger. Whereas subtracting 1 will take you closer to 0, and so because of that, the absolute value, the magnitude, or the distance from 0 will be less.

Now if you're on the other side, adding 1 will take you-- if you're on the other side, adding 1 will take you closer to 0. And so this magnitude would be smaller. Whereas, subtracting will take you farther away, so the absolute value actually increased the magnitude. And so this is the same as z being positive. And so this is the same as z being negative.

So what this tells you is that, if z is positive, then this probability is equal to 1. And if z is negative, this probability is equal to 0. Now why does that make sense? Well, it's because when lambda goes to infinity, you have the other case. Essentially, you pull this all the way up, really, really far, and it drops off really quickly.

And so when you take the limit, as lambda goes to infinity, effectively, it just becomes a spike at 0. And so, more or less, you're sure that y is going to be equal to 0. And so, effectively, z is actually going to be equal to x , effectively.

And because of that, because x can only be 1 or negative 1, then, depending on if you get a z that's positive, then you know for sure that it must have been that x was equal to 1. And if you get a z that's negative, you know for sure that it must have been that x was equal to negative 1. And so because of that, you get this interpretation.

And so we've looked at four different cases of the parameters. And in all four cases, our answer seems to make sense. And so we feel more confident in the answer.

And so to summarize, this whole problem involved using Bayes' rule. You start out with some distributions, and you apply Bayes' rule. And you go through the steps, and you plug-in the right terms. And then, in the end, it's always helpful to try to check your answers to make sure that it makes sense in some of the limiting.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013

Transcript – Lecture 10

The following content is provided under a Creative Commons license. Your support will help MIT OpenCourseWare continue to offer high quality educational resources for free. To make a donation or view additional materials from hundreds of MIT courses, visit MIT OpenCourseWare at ocw.mit.edu.

PROFESSOR: So today's agenda is to say a few more things about continuous random variables. Mainly we're going to talk a little bit about inference. This is a topic that we're going to revisit at the end of the semester. But there's a few things that we can already say at this point.

And then the new topic for today is the subject of derived distributions. Basically if you know the distribution of one random variable, and you have a function of that random variable, how to find a distribution for that function.

And it's a fairly mechanical skill, but that's an important one, so we're going to go through it. So let's see where we stand. Here is the big picture. That's all we have done so far. We have talked about discrete random variables, which we described by probability mass function. So if we have multiple random variables, we describe them with the a joint mass function.

And then we define conditional probabilities, or conditional PMFs, and the three are related according to this formula, which is, you can think of it either as the definition of conditional probability. Or as the multiplication rule, the probability of two things happening is the product of the probabilities of the first thing happening, and then the second happening, given that the first has happened.

There's another relation between this, which is the probability of x occurring, is the sum of the different probabilities of the different ways that x may occur, which is in conjunction with different values of y . And there's an analog of all that in the continuous world, where all you do is to replace p 's by f 's, and replace sums by integrals. So the formulas all look the same. The interpretations are a little more subtle, so the f 's are not probabilities, they're probability densities. So they're probabilities per unit length, or in the case of joint PDF's, these are probabilities per unit area. So they're densities of some sort.

Probably the more subtle concept to understand what it really is the conditional density. In some sense, it's simple. It's just the density of X in a world where you have been told the value of the random variable Y . It's a function that has two arguments, but the best way to think about it is to say that we fixed y . We're told the value of the random variable Y , and we look at it as a function of x . So as a function of x , the denominator is a constant, and it just looks like the joint density. when we keep y fixed. So it's really a function of one argument, just the argument x . And it has the same shape as the joint's density when you take that slice of it.

So conditional PDFs are just slices of joint PDFs.

There's a bunch of concepts, expectations, variances, cumulative distribution functions that apply equally well for both universes of discrete or continuous random variables. So why is probability useful? Probability is useful because, among other things, we use it to make sense of the world around us. We use it to make inferences about things that we do not see directly. And this is done in a very simple manner using the base rule. We've already seen some of that, and now we're going to revisit it with a bunch of different variations.

And the variations come because sometimes our random variable are discrete, sometimes they're continuous, or we can have a combination of the two. So the big picture is that there's some unknown random variable out there, and we know the distribution that's random variable. And in the discrete case, it's going to be given by PMF. In the continuous case, it's given a PDF. Then we have some phenomenon, some noisy phenomenon or some measuring device, and that measuring device produces observable random variables Y.

We don't know what x is, but we have some beliefs about how X is distributed. We observe the random variable Y. We need a model of this box. And the model of that box is going to be either a PMF, for the random variable Y. And that model tells us, if the true state of the world is X, how do we expect Y to be distributed? That's for the case where Y is this discrete. If Y is a continuous, you might instead have a density for Y, or something of that form.

So in either case, this should be a function that's known to us. This is our model of the measuring device. And now having observed y, we want to make inferences about x. What does it mean to make inferences? Well the most complete answer in the inference problem is to tell me the probability distribution of the unknown quantity.

But when I say the probability distribution, I don't mean this one. I mean the probability distribution that takes into account the measurements that you got. So the output of an inference problem is to come up with the distribution of X, the unknown quantity, given what we have already observed. And in the discrete case, it would be an object like that. If X is continuous, it would be an object of this kind.

OK, so we're given conditional probabilities of this type, and we want to get conditional distributions of the opposite type where the order of the conditioning is being reversed. So the starting point is always a formula such as this one. The probability of x happening, and then y happening given that x happens. This is the probability that a particular x and y happen simultaneously.

But this is also equal to the probability that y happens, and then that x happens, given that y has happened. And you take this expression and send one term to the denominator of the other side, and this gives us the base rule for the discrete case. Which is this one that you have already seen, and you have played with it.

So this is what the formula looks like in the discrete case. And the typical example where both random variables are discrete is the one we discussed some time ago. X is, let's say, a binary variable, or whether an airplane is present up there or not. Y is a discrete measurement, for example, whether our radar beeped or it didn't beep. And we make inferences and calculate the

probability that the plane is there, or the probability that the plane is not there, given the measurement that we have made.

And of course X and Y do not need to be just binary. They could be more general discrete random variables. So how does the story change in the continuous case? First, what's a possible application of the continuous case? Well, think of X as being some signal that takes values over a continuous range. Let's say X is the current through a resistor. And then you have some measuring device that measures currents, but that device is noisy, it gets hit, let's say for example, by Gaussian noise.

And the Y that you observe is a noisy version of X. But your instruments are analog, so you measure things on a continuous scale. What are you going to do in that case? Well the inference problem, the output of the inference problem, is going to be the conditional distribution of X. What do you think your current is based on a particular value of Y that you have observed?

So the output of our inference problem is, given the specific value of Y, to calculate this entire function as a function of x, and then go and plot it. How do we calculate it? You go through the same calculation as in the discrete case, except that all of the x's gets replaced by p's. In the continuous case, it's equally true that the joint's density is the product of the marginal density with the conditional density. So the formula is still valid with just a little change of notation. So we end up with the same formula here, except that we replace x's with p's.

So all of these functions are known to us. We have formulas for them. We fix a specific value of y, we plug it in, so we're left with a function of x. And that gives us the posterior distribution. Actually there's also a denominator term that's not necessarily given to us, but we can always calculate it if we have the marginal of X, and we have the model for measuring device. Then we can always find the marginal distribution of Y. So this quantity, that number, is in general a known one, as well, and doesn't give us any problems.

So to complicate things a little bit, we can also look into situations where our two random variables are of different kinds. For example, one random variable could be discrete, and the other it might be continuous. And there's two versions. Here one version is when X is discrete, but Y is continuous. What's an example of this?

Well suppose that I send a single bit of information so my X is 0 or 1. And what I measure is Y, which is X plus, let's say, Gaussian noise. This is the standard example that shows up in any textbook on communication, or signal processing. You send a single bit, but what you observe is a noisy version of that bit.

You start with a model of your x's. These would be your prior probabilities. For example, you might believe that either 0 or 1 are equally likely, in which case your PMF gives equal weight to two possible values. And then we need a model of our measuring device. This is one specific model. The general model would have a shape such as follows. Y has a distribution, its density. And that density, however, depends on the value of X.

So when x is 0, we might get a density of this kind. And when x is 1, we might get the density of a different kind. So these are the conditional densities of y in a universe that's specified by a particular value of x .

And then we go ahead and do our inference. OK, what's the right formula for doing this inference? We need a formula that's sort of an analog of this one, but applies to the case where we have two random variables of different kinds. So let me just redo this calculation here. Except that I'm not going to have a probability of taking specific values. It will have to be something a little different. So here's how it goes.

Let's look at the probability that X takes a specific value that makes sense in the discrete case, but for the continuous random variable, let's look at the probability that it takes values in some little interval. And now this probability of two things happening, I'm going to write it as a product. And I'm going to write this as a product in two different ways. So one way is to say that this is the probability that X takes that value and then given that X takes that value, the probability that Y falls inside that interval.

So this is our usual multiplication rule for multiplying probabilities, but I can use the multiplication rule also in a different way. It's the probability that Y falls in the range of interest. And then the probability that X takes the value of interest given that Y satisfies the first condition. So this is something that's definitely true. We're just using the multiplication rule. And now let's translate it into PMF is PDF notation.

So the entry up there is the PMF of X evaluated at x . The second entry, what is it? Well probabilities of little intervals are given to us by densities. But we are in the conditional universe where X takes on a particular value. So it's going to be the density of Y given the value of X times delta. So probabilities of little intervals are given by the density times the length of the little interval, but because we're working in the conditional universe, it has to be the conditional density.

Now let's try the second expression. This is the probability that the Y falls into the little interval. So that's the density of Y times delta. And then here we have an object which is the conditional probability X in a universe where the value of Y is given to us.

Now this relation is sort of approximate. This is true for very small delta in the limit. But we can cancel the deltas from both sides, and we're left with a formula that links together PMFs and PDFs. Now this may look terribly confusing because there's both p 's and f 's involved. But the logic should be clear. If a random variable is discrete, it's described by PMF. So here we're talking about the PMF of X in some particular universe. X is discrete, so it has a PMF.

Similarly here. Y is continuous so it's described by a PDF. And even in the conditional universe where I tell you the value of X , Y is still a continuous random variable, so it's been described by a PDF. So this is the basic relation that links together PMF and PDFs. In this mixed the world. And now in this inequality, you can take this term and send it to the new denominator to the other side. And what you end up with is the formula that we have up here.

And this is a formula that we can use to make inferences about the discrete random variable X when we're told the value of the continuous random variable Y. The probability that X takes on a particular value has something to do with the prior. And other than that, it's proportional to this quantity, the conditional of Y given X. So these are the quantities that we plotted here.

Suppose that the x's are equally likely in your prior, so we don't really care about that term. It tells us that the posterior of X is proportional to that particular density under the given x's. So in this picture, if I were to get a particular y here, I would say that x equals 1 has a probability that's proportional to this quantity. x equals 0 has a probability that's proportional to this quantity.

So the ratio of these two quantities gives us the relative odds of the different x's given the y that we have observed.

So we're going to come back to this topic and redo plenty of examples of these kinds towards the end of the class, when we spend some time dedicated to inference problems. But already at this stage, we sort of have the basic skills to deal with a lot of that. And it's useful at this point to pull all the formulas together.

So finally let's look at the last case that's remaining. Here we have a continuous phenomenon that we're trying to measure, but our measurements are discrete. What's an example where this might happen?

So you have some device that emits light, and you drive it with a current that has a certain intensity. You don't know what that current is, and it's a continuous random variable. But the device emits light by sending out individual photons. And your measurement is some other device that counts how many photons did you get in a single second.

So if we have devices that emit a very low intensity you can actually start counting individual photons as they're being observed. So we have a discrete measurement, which is the number of photons, and we have a continuous hidden random variable that we're trying to estimate. What do we do in this case?

Well we start again with a formula of this kind, and send the p term to the denominator. And that's the formula that we use there, except that the roles of x's and y's are interchanged. So since here we have Y being discrete, we should change all the subscripts. It would be $p_Y f_X$ given $y f_X$, and $P(Y \text{ given } X)$. So just change all those subscripts. Because now what we're used to be continuous became discrete, and vice versa.

Take that formula, send the other terms to the denominator, and we have a formula for the density, or X, given the particular measurements for Y that we have obtained.

In some sense that's all there is in Bayesian inference. It's using these very simple one line formulas. But why are there people then who make their living solving inference problems? Well, the devil is in the details. As we're going to discuss, there are some real world issues of how exactly do you design your f's, how do you model your system, then how do you do your calculations.

This might not be always easy. For example, there's certain integrals or sums that have to be evaluated, which may be hard to do and so on. So this object is a lot of richer than just these formulas. On the other hand, at the conceptual level, that's the basis for Bayesian inference, that these are the basic concepts.

All right, so now let's change gear and move to the new subject, which is the topic of finding the distribution of a functional for a random variable. We call those distributions derived distributions, because we're given the distribution of X . We're interested in a function of X . We want to derive the distribution of that function based on the distribution that we already know.

So it could be a function of just one random variable. It could be a function of several random variables. So one example that we are going to solve at some point, let's say you have to run the variables X and Y . Somebody tells you their distribution, for example, is a uniform of the square. For some reason, you're interested in the ratio of these two random variables, and you want to find the distribution of that ratio.

You can think of lots of cases where your random variable of interest is created by taking some other unknown variables and taking a function of them. And so it's legitimate to care about the distribution of that random variable.

A caveat, however. There's an important case where you don't need to find the distribution of that random variable. And this is when you want to calculate the expectations. If all you care about is the expected value of this function of the random variables, you can work directly with the distribution of the original random variables without ever having to find the PDF of g .

So you don't do unnecessary work if it's not needed, but if it's needed, or if you're asked to do it, then you just do it.

So how do we find the distribution of the function? As a warm-up, let's look at the discrete case. Suppose that X is a discrete random variable and takes certain values. We have a function g that maps x 's into y 's. And we want to find the probability mass function for Y .

So for example, if I'm interested in finding the probability that Y takes on this particular value, how would they find it? Well I ask, what are the different ways that these particular y value can happen? And the different ways that it can happen is either if x takes this value, or if X takes that value. So we identify this event in the y space with that event in the x space. These two events are identical. X falls in this set if and only if Y falls in that set.

Therefore, the probability of Y falling in that set is the probability of X falling in that set. The probability of X falling in that set is just the sum of the individual probabilities of the x 's in this set. So we just add the probabilities of the different x 's where the summation is taken over all x 's that leads to that particular value of y .

Very good. So that's all there is in the discrete case. It's a very nice and simple. So let's transfer these methods to the continuous case.

Suppose we are in the continuous case. Suppose that X and Y now can take values anywhere. And I try to use same methods and I ask, what is the probability that Y is going to take this value? At least if the diagram is this way, you would say this is the same as the probability that X takes this value. So I can find the probability of Y being this in terms of the probability of X being that.

Is this useful? In the continuous case, it's not. Because in the continuous case, any single value has 0 probability. So what you're going to get out of this argument is that the probability Y takes this value is 0, is equal to the probability that X takes that value which also 0.

That doesn't help us. We want to do something more. We want to actually find, perhaps, the density of Y , as opposed to the probabilities of individual y 's. So to find the density of Y , you might argue as follows. I'm looking at an interval for y , and I ask what's the probability of falling in this interval. And you go back and find the corresponding set of x 's that leads to those y 's, and equate those two probabilities.

The probability of all of those y 's collectively should be equal to the probability of all of the x 's that map into that interval collectively. And this way you can relate the two.

As far as the mechanics go, in many cases it's easier to not to work with little intervals, but instead to work with cumulative distribution functions that used to work with sort of big intervals. So you can instead do a different picture. Look at this set of y 's. This is the set of y 's that are smaller than a certain value. The probability of this set is given by the cumulative distribution of the random variable Y .

Now this set of y 's gets produced by some corresponding set of x 's. Maybe these are the x 's that map into y 's in that set. And then we argue as follows. The probability that the Y falls in this interval is the same as the probability that X falls in that interval. So the event of Y falling here and the event of X falling there are the same, so their probabilities must be equal. And then I do the calculations here. And I end up getting the cumulative distribution function of Y . Once I have the cumulative, I can get the density by just differentiating.

So this is the general cookbook procedure that we will be using to calculate it derived distributions.

We're interested in a random variable Y , which is a function of the x 's. We will aim at obtaining the cumulative distribution of Y . Somehow, manage to calculate the probability of this event. Once we get it, and what I mean by get it, I don't mean getting it for a single value of little y . You need to get this for all little y 's. So you need to get the function itself, the cumulative distribution. Once you get it in that form, then you can calculate the derivative at any particular point. And this is going to give you the density of Y .

So a simple two-step procedure. The devil is in the details of how you carry the mechanics. So let's do one first example. Suppose that X is a uniform random variable, takes values between 0 and 2. We're interested in the random variable Y , which is the cube of X . What kind of distribution is it going to have?

Now first notice that Y takes values between 0 and 8. So X is uniform, so all the x 's are equally likely. You might then say, well, in that case, all the y 's should be equally likely. So Y might also have a uniform distribution. Is this true? We'll find out.

So let's start applying the cookbook procedure. We want to find first the cumulative distribution of the random variable Y , which by definition is the probability that the random variable is less than or equal to a certain number. That's what we want to find. What we have in our hands is the distribution of X . That's what we need to work with. So the first step that you need to do is to look at this events and translate it, and write it in terms of the random variable about which you know you have information.

So Y is X cubed, so this event is the same as that event. So now we can forget about the y 's. It's just an exercise involving a single random variable with a known distribution and we want to calculate the probability of some event.

So we're looking at this event. X cubed being less than or equal to Y . We massage that expression so that's it involves X directly, so let's take cubic roots of both sides of this inequality. This event is the same as the event that X is less than or equal to Y to the $1/3$. Now with a uniform distribution on $[0,2]$, what is that probability going to be?

It's the probability of being in the interval from 0 to y to the $1/3$, so it's going to be in the area under the uniform going up to that point. And what's the area under that uniform?

So here's x . Here is the distribution of X . It goes up to 2. The distribution of X is this one. We want to go up to y to the $1/3$. So the probability for this event happening is this area. And the area is equal to the base, which is y to the $1/3$ times the height. What is the height?

Well since the density must integrate to 1, the total area under the curve has to be 1. So the height here is $1/2$, and that explains why we get the $1/2$ factor down there.

So that's the formula for the cumulative distribution. And then the rest is easy. You just take derivatives. You differentiate this expression with respect to y $1/2$ times $1/3$, and y drops by one power. So you get y to $2/3$ in the denominator.

So if you wish to plot this, it's $1/y$ to the $2/3$. So when y goes to 0, it sort of blows up and it goes on this way. Is this picture correct the way I've drawn it? What's wrong with it?

[? AUDIENCE: Something. ?]

PROFESSOR: Yes. y only takes values from 0 to 8. This formula that I wrote here is only correct when the previous picture applies. I took my y to the $1/3$ to be between 0 and 2. So this formula here is only correct for y between 0 and 8. And for that reason, the formula for the derivative is also true only for a y between 0 and 8. And any other values of y are impossible, so they get zero density. So to complete the picture here, the PDF of y has a cut-off of 8, and it's also 0 everywhere else.

And one thing that we see is that the distribution of Y is not uniform. Certain y 's are more likely than others, even though we started with a uniform random variable X .

All right. So we will keep doing examples of this kind, a sequence of progressively more interesting or more complicated. So that's going to continue in the next lecture. You're going to see plenty of examples in your recitations and tutorials and so on. So let's do one that's pretty similar to the one that we did, but it's going to add to just a small twist in how we do the mechanics.

OK so you set your cruise control when you start driving. And you keep driving at the constants based at the constant speed. Where you set your cruise control is somewhere between 30 and 60. You're going to drive a distance of 200. And so the time it's going to take for your trip is 200 over the setting of your cruise control. So it's $200/V$.

Somebody gives you the distribution of V , and they tell you not only it's between 30 and 60, it's roughly equally likely to be anything between 30 and 60, so we have a uniform distribution over that range. So we have a distribution of V . We want to find the distribution of the random variable T , which is the time it takes till your trip ends.

So how are we going to proceed? We'll use the exact same cookbook procedure. We're going to start by finding the cumulative distribution of T . What is this? By definition, the cumulative distribution is the probability that T is less than a certain number. OK. Now we don't know the distribution of T , so we cannot work with these event directly. But we take that event and translate it into T -space. So we replace the t 's by what we know T to be in terms of V or the v 's

All right. So we have the distribution of V . So now let's calculate this quantity. OK. Let's massage this event and rewrite it as the probability that V is larger or equal to $200/T$.

So what is this going to be? So let's say that $200/T$ is some number that falls inside the range. So that's going to be true if $200/T$ is bigger than 30, and less than 60. Which means that t is less than $30/200$. No, $200/30$. And bigger than $200/60$. So for t 's inside that range, this number $200/t$ falls inside that range. This is the range of t 's that are possible, given the description of the problem the we have set up.

So for t 's in that range, what is the probability that V is bigger than this number? So V being bigger than that number is the probability of this event, so it's going to be the area under this curve. So the area under that curve is the height of the curve, which is $1/3$ over 30 times the base. How big is the base? Well it's from that point to 60, so the base has a length of $60 - 200/t$.

And this is a formula which is valid for those t 's for which this picture is correct. And this picture is correct if $200/T$ happens to fall in this interval, which is the same as T falling in that interval, which are the t 's that are possible.

So finally let's find the density of T , which is what we're looking for. We find this by taking the derivative in this expression with respect to t . We only get one term from here. And this is going to be $200/30$, 1 over t squared.

And this is the formula for the density for t's in the allowed to range. OK, so that's the end of the solution to this particular problem as well. I said that there was a little twist compared to the previous one. What was the twist? Well the twist was that in the previous problem we dealt with the X cubed function, which was monotonically increasing. Here we dealt with the function that was monotonically decreasing. So when we had to find the probability that T is less than something, that translated into an event that V was bigger than something. Your time is less than something if and only if your velocity is bigger than something.

So for when you're dealing with the monotonically decreasing function, at some point some inequalities will have to get reversed.

Finally let's look at a very useful one. Which is the case where we take a linear function of a random variable. So X is a random variable with given distribution, and we can see there is a linear function. So in this particular instance, we take a to be equal to 2 and b equal to 5. And let us first argue just by picture.

So X is a random variable that has a given distribution. Let's say it's this weird shape here. And x ranges from -1 to +2. Let's do things one step at the time. Let's first find the distribution of $2X$. Why do you think you know about $2X$? Well if x ranges from -1 to 2, then the random variable X is going to range from -2 to +4. So that's what the range is going to be.

Now dealing with the random variable $2X$, as opposed to the random variable X , in some sense it's just changing the units in which we measure that random variable. It's just changing the scale on which we draw and plot things. So if it's just a scale change, then intuition should tell you that the random variable X should have a PDF of the same shape, except that it's scaled out by a factor of 2, because our random variable of $2X$ now has a range that's twice as large.

So we take the same PDF and scale it up by stretching the x-axis by a factor of 2. So what does scaling correspond to in terms of a formula? So the distribution of $2X$ as a function, let's say, a generic argument z , is going to be the distribution of X , but scaled by a factor of 2.

So taking a function and replacing its arguments by the argument over 2, what it does is it stretches it by a factor of 2. You have probably been tortured ever since middle school to figure out when need to stretch a function, whether you need to put $2z$ or $z/2$. And the one that actually does the stretching is to put the $z/2$ in that place. So that's what the stretching does.

Could that to be the full answer? Well there's a catch. If you stretch this function by a factor of 2, what happens to the area under the function? It's going to get doubled. But the total probability must add up to 1, so we need to do something else to make sure that the area under the curve stays to 1. So we need to take that function and scale it down by this factor of 2.

So when you're dealing with a multiple of a random variable, what happens to the PDF is you stretch it according to the multiple, and then scale it down by the same number so that you preserve the area under that curve. So now we found the distribution of $2X$.

How about the distribution of $2X + 5$? Well what does adding 5 to random variable do? You're going to get essentially the same values with the same probability, except that those values all get shifted by 5. So all that you need to do is to take this PDF here, and shift it by 5 units. So the range used to be from -2 to 4. The new range is going to be from 3 to 9. And that's the final answer. This is the distribution of $2X + 5$, starting with this particular distribution of X .

Now shifting to the right by b , what does it do to a function? Shifting to the right to by a certain amount, mathematically, it corresponds to putting $-b$ in the argument of the function. So I'm taking the formula that I had here, which is the scaling by a factor of a . The scaling down to keep the total area equal to 1. And then I need to introduce this extra term to do the shifting.

So this is a plausible argument. The proof by picture that this should be the right answer. But just in order to keep our skills tuned and refined, let us do this derivation in a more formal way using our two-step cookbook procedure. And I'm going to do it under the assumption that a is positive, as in the example that's we just did.

So what's the two-step procedure? We want to find the cumulative of Y , and after that we're going to differentiate. By definition the cumulative is the probability that the random variable takes values less than a certain number. And now we need to take this event and translate it, and express it in terms of the original random variables.

So Y is, by definition, $aX + b$, so we're looking at this event. And now we want to express this event in a clean form where X shows up in a straight way. Let's say I'm going to massage this event and write it in this form. For this inequality to be true, x should be less than or equal to $(y - b)$ divided by a .

OK, now what is this? This is the cumulative distribution of X evaluated at the particular point. So we got a formula for the cumulative Y based on the cumulative of X . What's the next step? Next step is to take derivatives of both sides. So the density of Y is going to be the derivative of this expression with respect to y . OK, so now here we need to use the chain rule. It's going to be the derivative of the F function with respect to its argument. And then we need to take the derivative of the argument with respect to y .

What is the derivative of the cumulative? The derivative of the cumulative is the density itself. And we evaluate it at the point of interest. And then the chain rule tells us that we need to take the derivative of this with respect to y , and the derivative of this with respect to y is $1/a$. And this gives us the formula which is consistent with what I had written down here, for the case where a is a positive number.

What if a was a negative number? Could this formula be true? Of course not. Densities cannot be negative, right? So that formula cannot be true. Something needs to change. What should change? Where does this argument break down when a is negative?

So when I write this inequality in this form, I divide by a . But when you divide by a negative number, the direction of an inequality is going to change. So when a is negative, this inequality becomes larger than or equal to. And in that case, the expression that I have up there would

change when this is larger than here. Instead of getting the cumulative, I would get 1 minus the cumulative of $(y - b)$ divided by a .

So this is the probability that X is bigger than this particular number. And now when you take the derivatives, there's going to be a minus sign that shows up. And that minus sign will end up being here. And so we're taking the negative of a negative number, and that basically is equivalent to taking the absolute value of that number.

So all that happens when we have a negative a is that we have to take the absolute value of the scaling factor instead of the factor itself.

All right, so this general formula is quite useful for dealing with linear functions of random variables. And one nice application of it is to take the formula for a normal random variable, consider a linear function of a normal random variable, plug into this formula, and what you will find is that Y also has a normal distribution. So using this formula, now we can prove a statement that I had made a couple of lectures ago, that a linear function of a normal random variable is also linear. That's how you would prove it. I think this is it for today so.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013 Transcript – Tutorial: The Probability Distribution Function (PDF) of [X]

Hi, In this problem, we'll be looking at the PDF the absolute value of x . So if we know a random variable, x , and we know its PDF, how can we use that information to help us find the PDF of another random variable-- the absolute value of x ? And so throughout this problem, we'll define a new random variable called y . And we'll define that y to be equal to the absolute value of x , just to make things simpler.

So we'll do a couple of concrete examples, and then we'll try to generalize at the end. The first example that we'll deal with in part A is this PDF for x . So we're told that the PDF of x is $1/3$ between negative 2 and 1, and 0 otherwise. And here's a picture of what it looks like. It's just a rectangle from negative 2 to 1. So now we want to find out what is the PDF of the absolute value of x , which we've called y ?

And at this point, it may be helpful to step back and think about this problem from the discrete point of view again. So if x were a discrete random variable, the problem would be, what is the probability that the absolute value of x is equal to, say, $1/2$? Well, the probability that the absolute value of x is equal to $1/2$ -- that can occur in two different ways. One is that x itself is $1/2$. Or x could be negative $1/2$, in which case, the absolute value of x would still be $1/2$.

So those two events are mutually exclusive. And so the probability of either one of them happening is you can just add them up. And so the probability of the absolute value of x being $1/2$ would have two contributions, one from x being $1/2$, and one from x being negative $1/2$.

The analogous idea carries over to the continuous case, when you have a PDF. So now let's say that we're interested in the case where we want to know the PDF of y at $1/2$. Well, that again, is going to have two contributions, one from where x is $1/2$, and one from where x is minus $1/2$. And so you can just imagine that each one of these values for y -- and remember, y has to be non-negative, because it's an absolute value-- has two contributions, one from the right side of 0, and one from the left, or negative, side of 0.

So you can come up and write an algebraic expression for this, and we'll do that in Part C. But you can also look at this from a visual point of view. And you can take the PDF diagram itself and imagine transforming it to find out what the PDF of the absolute value of x would look like. So the way to do it would be you take what's on the negative side. You flip it over and take the mirror image, and then you stack it on top of what you have on the right-hand side, or the positive side.

So take this, flip it over, and stack it on top. You can imagine just taking this block, flipping it over. And just think of it as like a Tetris block that's falling down from above. And it stacks on top of wherever it lands. So it'll turn it into something that looks like this.

So there's already a block of height $1/3$ from 0 to 1. That's from the original 1. And now we take this, and flip it over, and drop it on top. Well, this part is going to fall on top of the segment from 0 to 1.

And then this part gets flipped over and dropped over here. And it falls down here. And so the final PDF actually looks like this kind of staircase, where this is $2/3$ now, because this has two contributions of $1/3$ each, and this is $1/3$.

So that is the graphical way of approaching this. And the PDF for completeness, the PDF of y would be $2/3$ for y between 0 and 1, $1/3$ for y from 1 to 2, and 0 otherwise.

All right, so let's move on to part B, and get some more practice. Part B, we're given that this PDF of x now is 2 times e to the negative $2x$ for x positive, and 0 otherwise. Now you may just recognize this as an exponential random variable with a parameter of 2. And again, we can graph this and see what it looks like. And it turns out that it's going to start out at 2 and fall off exponentially.

So in this case, this is actually quite simple. Because if you look at it, x is already positive. It doesn't have any negative parts. So in fact, the absolute value of x is the same as x itself, because x is never negative.

And so y is just the same thing as x . And so in this case, actually, the PDF of y is exactly the same as the PDF of x . It's just $2e$ to the minus $2y$, for y positive and zero otherwise.

Now you can also see this graphically also, because to the left of 0, the negative part, there is no PDF. The PDF is 0. And so if you were to take this, flip it over, and drop it on top, you wouldn't get anything, because there's nothing there. And so the entire PDF, even after you take the absolute value, is just the original one.

So to generalize, what I said at the beginning was that, remember, the probability in the discrete case, if you wanted the probability that the absolute value of a random variable equals something, that would just be the probability that the random variable equals that value of little x , or the random variable equals negative little x .

In either of those two cases, the absolute value would equal x . So you get those two contributions. And so to generalize in the continuous case with PDFs, you get something that looks very similar. So in this case, the PDF of y is just the PDF of x at y .

So this is the case where x is just equal to y , plus the PDF of x evaluated negative y . So you, again, have both of these two contributions. And we can rewrite this top one to make it look more similar. So the PMF of some discrete [? number ?] y , where this is a discrete random variable that's equal to the absolute value of x , would be the PMF of x evaluated at y , plus the PMF of x evaluated at negative y . So in both the discrete and continuous cases, you have the same thing.

So the overall summary of this problem is that, when you take a transformation-- in this case, an absolute value-- you can reason about it and figure out how to decompose that into arguments about the original random variable, just plain old x . And for the specific case of the absolute value, it just becomes taking a mirror image and popping it on top of what you originally had. So remember, you always have these two contributions.

And so if you ever have a random variable that you need to take an absolute value of, you don't have to be scared. All you have to do is consider both of these contributions and add them up, and you have the PDF that you want. So I'll see you next time.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

LECTURE 10

Continuous Bayes rule; Derived distributions

- **Readings:**
Section 3.6; start Section 4.1

Review

$$\begin{aligned} p_X(x) &= f_X(x) \\ p_{X,Y}(x,y) &= f_{X,Y}(x,y) \\ p_{X|Y}(x|y) &= \frac{p_{X,Y}(x,y)}{f_Y(y)} = \frac{f_{X,Y}(x,y)}{f_Y(y)} \\ p_X(x) &= \sum_y p_{X,Y}(x,y) \quad f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) dy \end{aligned}$$

$$F_X(x) = P(X \leq x)$$

$$E[X], \text{ var}(X)$$

The Bayes variations

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)} = \frac{p_X(x)p_{Y|X}(y|x)}{p_Y(y)}$$

$$p_Y(y) = \sum_x p_X(x)p_{Y|X}(y|x)$$

Example:

- $X = 1, 0$: airplane present/not present
- $Y = 1, 0$: something did/did not register on radar

Continuous counterpart

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{f_X(x)f_{Y|X}(y|x)}{f_Y(y)}$$

$$f_Y(y) = \int_x f_X(x)f_{Y|X}(y|x) dx$$

- Example:** X : some signal; “prior” $f_X(x)$
 Y : noisy version of X
 $f_{Y|X}(y|x)$: model of the noise

Discrete X , Continuous Y

$$p_{X|Y}(x|y) = \frac{p_X(x)f_{Y|X}(y|x)}{f_Y(y)}$$

$$f_Y(y) = \sum_x p_X(x)f_{Y|X}(y|x)$$

Example:

- X : a discrete signal; “prior” $p_X(x)$
- Y : noisy version of X
- $f_{Y|X}(y|x)$: continuous noise model

Continuous X , Discrete Y

$$f_{X|Y}(x|y) = \frac{f_X(x)p_{Y|X}(y|x)}{p_Y(y)}$$

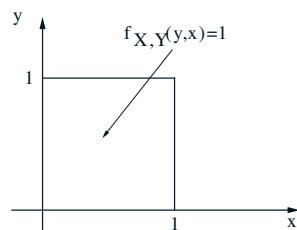
$$p_Y(y) = \int_x f_X(x)p_{Y|X}(y|x) dx$$

Example:

- X : a continuous signal; “prior” $f_X(x)$ (e.g., intensity of light beam);
- Y : discrete r.v. affected by X (e.g., photon count)
- $p_{Y|X}(y|x)$: model of the discrete r.v.

What is a derived distribution

- It is a PMF or PDF of a function of one or more random variables with known probability law. E.g.:



- Obtaining the PDF for

$$g(X, Y) = Y/X$$

involves deriving a distribution.

Note: $g(X, Y)$ is a random variable

When not to find them

- Don’t need PDF for $g(X, Y)$ if only want to compute expected value:

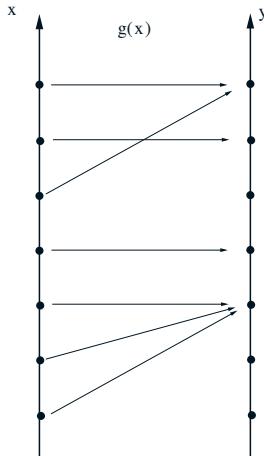
$$E[g(X, Y)] = \int \int g(x, y) f_{X,Y}(x, y) dx dy$$

How to find them

- **Discrete case**

- Obtain probability mass for each possible value of $Y = g(X)$

$$p_Y(y) = P(g(X) = y) = \sum_{x: g(x)=y} p_X(x)$$



The continuous case

- **Two-step procedure:**

- Get CDF of Y : $F_Y(y) = P(Y \leq y)$

- Differentiate to get

$$f_Y(y) = \frac{dF_Y}{dy}(y)$$

Example

- X : uniform on $[0,2]$

- Find PDF of $Y = X^3$

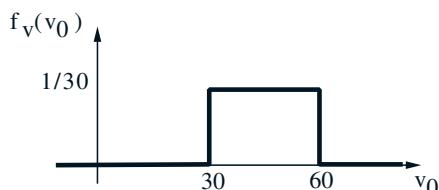
- **Solution:**

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(X^3 \leq y) \\ &= P(X \leq y^{1/3}) = \frac{1}{2}y^{1/3} \end{aligned}$$

$$f_Y(y) = \frac{dF_Y}{dy}(y) = \frac{1}{6y^{2/3}}$$

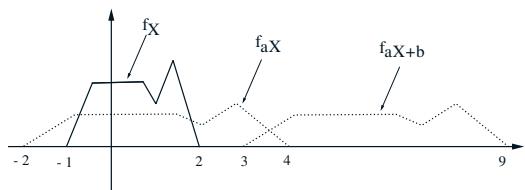
Example

- Joan is driving from Boston to New York. Her speed is uniformly distributed between 30 and 60 mph. What is the distribution of the duration of the trip?
- Let $T(V) = \frac{200}{V}$.
- Find $f_T(t)$



The pdf of $Y=aX+b$

$$Y = 2X + 5:$$



$$f_Y(y) = \frac{1}{|a|} f_X \left(\frac{y-b}{a} \right)$$

- Use this to check that if X is normal, then $Y = aX + b$ is also normal.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Recitation 11
October 14, 2010

- Let X be a discrete random variable that takes the values 1 with probability p and -1 with probability $1 - p$. Let Y be a continuous random variable independent of X with the Laplacian (two-sided exponential) distribution

$$f_Y(y) = \frac{1}{2}\lambda e^{-\lambda|y|},$$

and let $Z = X + Y$. Find $\mathbf{P}(X = 1 | Z = z)$. Check that the expression obtained makes sense for $p \rightarrow 0^+$, $p \rightarrow 1^-$, $\lambda \rightarrow 0^+$, and $\lambda \rightarrow \infty$.

- Let Q be a continuous random variable with PDF

$$f_Q(q) = \begin{cases} 6q(1-q), & \text{if } 0 \leq q \leq 1, \\ 0, & \text{otherwise.} \end{cases}$$

This Q represents the probability of success of a Bernoulli random variable X , i.e.,

$$\mathbf{P}(X = 1 | Q = q) = q.$$

Find $f_{Q|X}(q|x)$ for $x \in \{0, 1\}$ and all q .

- Let X have the normal distribution with mean 0 and variance 1, i.e.,

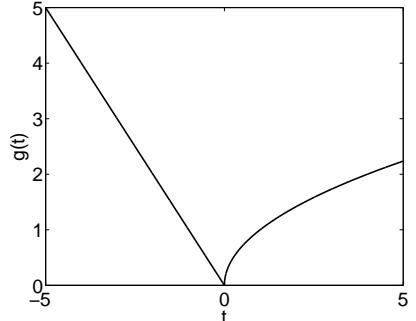
$$f_X(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}.$$

Also, let $Y = g(X)$ where

$$g(t) = \begin{cases} -t, & \text{for } t \leq 0; \\ \sqrt{t}, & \text{for } t > 0, \end{cases}$$

as shown to the right.

Find the probability density function of Y .



MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Recitation 11 Solutions
October 14, 2010

1. We need to apply the version of Bayes rule for a discrete random variable conditioned on a continuous random variable:

$$p_{X|Z}(x | z) = \frac{p_X(x)f_{Z|X}(z | x)}{f_Z(z)} = \frac{p_X(x)f_{Z|X}(z | x)}{\sum_{k=0}^1 p_X(k)f_{Z|X}(z | k)}.$$

Specifically,

$$\begin{aligned} \mathbf{P}(X = 1 | Z = z) &= p_{X|Z}(1 | z) = \frac{p_X(1)f_{Z|X}(z | 1)}{\sum_{k=0}^1 p_X(k)f_{Z|X}(z | k)} \\ &= \frac{p \frac{1}{2} \lambda e^{-\lambda|z-1|}}{(1-p) \frac{1}{2} \lambda e^{-\lambda|z+1|} + p \frac{1}{2} \lambda e^{-\lambda|z-1|}} \\ &= \frac{p e^{-\lambda|z-1|}}{(1-p)e^{-\lambda|z+1|} + p e^{-\lambda|z-1|}} \\ &= \frac{p e^{-\lambda|z-1|}}{(1-p)e^{-\lambda|z+1|} + p e^{-\lambda|z-1|}} \cdot \frac{e^{\lambda|z-1|}}{e^{\lambda|z-1|}} \\ &= \frac{p}{(1-p)e^{-\lambda(|z+1|-|z-1|)} + p} \end{aligned}$$

The final manipulations are to ease interpretations for $p \rightarrow 0^+$, $p \rightarrow 1^-$, $\lambda \rightarrow 0^+$, and $\lambda \rightarrow \infty$. Easily

$$\lim_{p \rightarrow 0^+} \mathbf{P}(X = 1 | Z = z) = 0 \quad \text{and} \quad \lim_{p \rightarrow 1^-} \mathbf{P}(X = 1 | Z = z) = 1;$$

these make sense because the observation z should become unimportant when value of X becomes certain without it. Next,

$$\lim_{\lambda \rightarrow 0^+} \mathbf{P}(X = 1 | Z = z) = p,$$

which makes sense because the distribution of Y becomes very flat as $\lambda \rightarrow 0^+$, making the observation uninformative. Finally,

$$\lim_{\lambda \rightarrow \infty} \mathbf{P}(X = 1 | Z = z) = \begin{cases} 1, & \text{if } |z+1| > |z-1|, \\ 0, & \text{if } |z+1| < |z-1|, \end{cases} = \begin{cases} 1, & \text{if } z > 0, \\ 0, & \text{if } z < 0; \end{cases}$$

this makes sense because $\lambda \rightarrow \infty$ makes the Y negligible.

2. We need to apply the version of Bayes rule for a continuous random variable conditioned on a discrete random variable:

$$f_{Q|X}(q | x) = \frac{f_Q(q)p_{X|Q}(x | q)}{p_X(x)} = \frac{f_Q(q)p_{X|Q}(x | q)}{\int_0^1 f_Q(q)p_{X|Q}(x | q) dq}.$$

For $x = 0$ and $q \in [0, 1]$,

$$\begin{aligned} f_{Q|X}(q | 0) &= \frac{f_Q(q)p_{X|Q}(0 | q)}{\int_0^1 f_Q(q)p_{X|Q}(0 | q) dq} = \frac{6q(1-q) \cdot (1-q)}{\int_0^1 6q(1-q)(1-q) dq} \\ &= \frac{6q(1-q) \cdot (1-q)}{1/2} = 12q(1-q)^2. \end{aligned}$$

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
 (Fall 2010)

For $x = 1$ and $q \in [0, 1]$,

$$\begin{aligned} f_{Q|X}(q | 1) &= \frac{f_Q(q)p_{X|Q}(1 | q)}{\int_0^1 f_Q(q)p_{X|Q}(1 | q) dq} = \frac{6q(1-q) \cdot q}{\int_0^1 6q(1-q)q dq} \\ &= \frac{6q(1-q) \cdot q}{1/2} = 12q^2(1-q). \end{aligned}$$

The distributions $f_Q(q)$, $f_{Q|X}(q | 0)$, and $f_{Q|X}(q | 1)$ are all in the family of *beta distributions*, which arise again in Chapter 8.

3. Because of the definition of g , the random variable Y takes on only nonnegative values. Thus $f_Y(y) = 0$ for any negative y . For $y > 0$,

$$\begin{aligned} F_Y(y) &= \mathbf{P}(Y \leq y) \\ &= \mathbf{P}(X \in [-y, 0]) + \mathbf{P}(X \in (0, y^2]) \\ &= (F_X(0) - F_X(-y)) + (F_X(y^2) - F_X(0)) \\ &= F_X(y^2) - F_X(-y). \end{aligned}$$

Taking the derivative of $F_Y(y)$ (and using the chain rule),

$$\begin{aligned} f_Y(y) &= 2yf_X(y^2) + f_X(-y) \\ &= \frac{1}{\sqrt{2\pi}} \left(2ye^{-y^4/2} + e^{-y^2/2} \right). \end{aligned}$$

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Tutorial 5
October 14/15, 2010

1. Let Q be a random variable which is uniformly distributed between 0 and 1. On any given day, a particular machine is functional with probability Q . Furthermore, given the value of Q , the status of the machine on different days is independent.

- Find the probability that the machine is functional on a particular day.
- We are told that the machine was functional on m out of the last n days. Find the conditional PDF of Q . You may use the identity

$$\int_0^1 p^k (1-p)^{n-k} dp = \frac{k!(n-k)!}{(n+1)!}$$

2. Let X be a random variable with PDF f_X . Find the PDF of the random variable $Y = |X|$

- when $f_X(x) = \begin{cases} 1/3, & \text{if } -2 < x \leq 1, \\ 0, & \text{otherwise;} \end{cases}$
 - when $f_X(x) = \begin{cases} 2e^{-2x}, & \text{if } x > 0, \\ 0, & \text{otherwise;} \end{cases}$
 - for general $f_X(x)$.
3. An ambulance travels back and forth, at a constant specific speed v , along a road of length ℓ . We may model the location of the ambulance at any moment in time to be uniformly distributed over the interval $(0, \ell)$. Also at any moment in time, an accident (not involving the ambulance itself) occurs at a point uniformly distributed on the road; that is, the accident's distance from one of the fixed ends of the road is also uniformly distributed over the interval $(0, \ell)$. Assume the location of the accident and the location of the ambulance are independent.

Supposing the ambulance is capable of *immediate* U-turns, compute the CDF and PDF of the ambulance's travel time T to the location of the accident.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Tutorial 5: Solutions

1. (a) Let A be the event that the machine is functional. Conditioned on the random variable Q taking on a particular value q , $\mathbf{P}(A|Q = q) = q$. Using the continuous form of the total probability theorem, the probability of event A is given by:

$$\begin{aligned}\mathbf{P}(A) &= \int_0^1 \mathbf{P}(A|Q = q) f_Q(q) dq \\ &= \int_0^1 q dq \\ &= 1/2\end{aligned}$$

- (b) Let B be the event that the machine is functional on m out of the last n days. Conditioned on random variable Q taking on value q (a probability q of being functional) the probability of event B is binomial with n trials, m successes, and a probability q of success in each trial. Again using the total probability theorem, the probability of event B is given by:

$$\begin{aligned}\mathbf{P}(B) &= \int_0^1 \mathbf{P}(B|Q = q) f_Q(q) dq \\ &= \int_0^1 \binom{n}{m} q^m (1-q)^{n-m} f_Q(q) dq \\ &= \binom{n}{m} \frac{m!(n-m)!}{(n+1)!}\end{aligned}$$

We then find the distribution on Q conditioned on event B using Bayes rule:

$$\begin{aligned}f_{Q|B}(q) &= \frac{\mathbf{P}(B|Q = q) f_Q(q)}{\mathbf{P}(B)} \\ &= \frac{q^m (1-q)^{n-m}}{\frac{m!(n-m)!}{(n+1)!}} \quad 0 \leq q \leq 1, \quad n \geq m.\end{aligned}$$

2. Since $Y = |X|$ you can visualize the PDF for any given y as

$$f_Y(y) = \begin{cases} f_X(y) + f_X(-y), & \text{if } y \geq 0, \\ 0, & \text{if } y < 0. \end{cases}$$

Also note that since $Y = |X|$, $Y \geq 0$.

$$(a) f_X(x) = \begin{cases} \frac{1}{3}, & \text{if } -2 < x \leq 1, \\ 0, & \text{otherwise.} \end{cases}$$

So, $f_X(x)$ for $-1 \leq x \leq 0$ gets added to $f_X(x)$ for $0 \leq x \leq 1$:

$$f_Y(y) = \begin{cases} 2/3, & \text{if } 0 \leq y \leq 1, \\ 1/3, & \text{if } 1 < y \leq 2, \\ 0, & \text{otherwise.} \end{cases}$$

- (b) Here we are told $X > 0$. So there are no negative values of X that need to be considered. Thus,

$$f_Y(y) = f_X(y) = \begin{cases} 2e^{-2y}, & \text{if } y \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

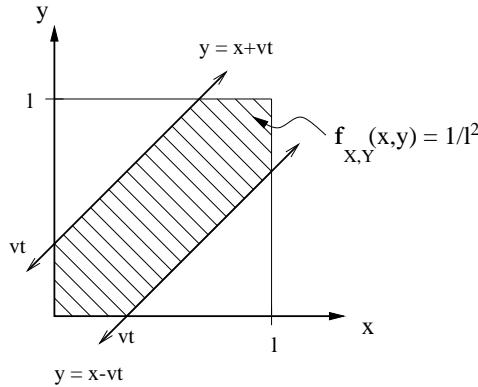
- (c) As explained in the beginning, $f_Y(y) = f_X(y) + f_X(-y)$.

3. We want to compute the CDF of the ambulance's travel time T , $\mathbf{P}(T \leq t) = \mathbf{P}(|X - Y| \leq vt)$, where X and Y are the locations of the ambulance and accident (uniform over $[0, l]$). Since X and Y are independent, we know:

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{l^2}, & \text{if } 0 \leq x, y \leq l \\ 0, & \text{otherwise} \end{cases}.$$

$$\begin{aligned} \mathbf{P}(T \leq t) &= \mathbf{P}(|X - Y| \leq vt) = \mathbf{P}(-vt \leq Y - X \leq vt) \\ &= \mathbf{P}(X - vt \leq Y \leq X + vt) \end{aligned}$$

We can see that $\mathbf{P}(X - vt \leq Y \leq X + vt)$ corresponds to the integral of the joint density of X and Y over the shaded region in the figure below:



Therefore, because the joint density is uniform over the entire region, we have:

$$F_T(t) = (1/l^2) \times (\text{Shaded area}) = \begin{cases} 0 & \text{if } t < 0 \\ \frac{2vt}{l} - \frac{(vt)^2}{l^2} & \text{if } 0 \leq t < \frac{l}{v} \\ 1 & \text{if } t \geq \frac{l}{v} \end{cases}.$$

By differentiating the CDF, we find the density of T :

$$f_T(t) = \begin{cases} \frac{2v}{l} - \frac{2v^2t}{l^2} & \text{if } 0 \leq t \leq \frac{l}{v} \\ 0 & \text{otherwise} \end{cases}.$$

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013 Transcript – Lecture 11

The following content is provided under a Creative Commons license. Your support will help MIT OpenCourseWare continue to offer high-quality, educational resources for free. To make a donation or view additional materials from hundreds of MIT courses, visit MIT OpenCourseWare at ocw.mit.edu.

PROFESSOR: Good morning. So today we're going to continue the subject from last time. So we're going to talk about derived distributions a little more, how to derive the distribution of a function of a random variable. So last time we discussed a couple of examples in which we had a function of a single variable. And we found the distribution of Y , if we're told the distribution of X .

So today we're going to do an example where we deal with the function of two random variables. And then we're going to consider the most interesting example of this kind, in which we have a random variable of the form W , which is the sum of two independent, random variables. That's a case that shows up quite often. And so we want to see what exactly happens in this particular case.

Just one comment that I should make. The material that we're covering now, chapter four, is sort of conceptually a little more difficult than one we have been doing before. So I would definitely encourage you to read the text before you jump and try to do the problems in your problem sets.

OK, so let's start with our example, in which we're given two random variables. They're jointly continuous. And their distribution is pretty simple. They're uniform on the unit square. In particular, each one of the random variables is uniform on the unit interval. And the two random variables are independent.

What we're going to find is the distribution of the ratio of the two random variables. How do we go about it? Well, the same cookbook procedure that we used last time for the case of a single random variable. The cookbook procedure that we used for this case also applies to the case where you have a function of multiple random variables.

So what was the cookbook procedure? The first step is to find the cumulative distribution function of the random variable of interest and then take the derivative in order to find the density. So let's find the cumulative. So, by definition, the cumulative is the probability that the random variable is less than or equal to the argument of the cumulative. So if we write this event in terms of the random variable of interest, this is the probability that our random variable is less than or equal to z .

So what is that? OK, so the ratio is going to be less than or equal to z , if and only if the pair, (x,y) , happens to fall below the line that has a slope z . OK, so we draw a line that has a slope z . The ratio is less than this number, if and only if we get the pair of x and y that falls inside this triangle.

So we're talking about the probability of this particular event. Since this line has a slope of z , the height at this point is equal to z . And so we can find the probability of this event.

It's just the area of this triangle. And so the area is 1 times z times $1/2$. And we get the answer, $z/2$.

Now, is this answer always correct? Now, this answer is going to be correct only if the slope happens to be such that we get a picture of this kind. So when do we get a picture of this kind? When the slope is less than 1.

If I consider a different slope, a number, little z -- that happens to be a slope of that kind -- then the picture changes. And in that case, we get a picture of this kind, let's say. So this is a line here of slope z , again. And this is the second case in which our number, little z , is bigger than 1.

So how do we proceed? Once more, the cumulative is the probability that the ratio is less than or equal to that number. So it's the probability that we fall below the red line.

So we're talking about the event, about this event. So to find the probability of this event, we need to find the area of this red shape. And one way of finding this area is to consider the whole area and subtract the area of this triangle.

So let's do it this way. It's going to be 1 minus the area of the triangle. Now, what's the area of the triangle? It's $1/2$ times this side, which is 1 times this side.

How big is that side? Well, if y and the slope is z , now z is the ratio y over x . So if y over x -- at this point we have $y/x = z$ and $y=1$. This means that z is $1/x$.

So the coordinate of this point is $1/x$. And this means that we're going to-- $1/z$ So here we get the factor of $1/z$.

And we're basically done. I guess if you want to have a complete answer, you should also give the formula for z less than 0. What is the cumulative when z is less than 0, the probability that you get the ratio that's negative?

Well, since our random variables are positive, there's no way that you can get a negative ratio. So the cumulative down there is equal to 0. So we can plot the cumulative. And we can take its derivative in order to find the density.

So the cumulative that we got starts at 0, when z 's are negative. Then it starts going up in proportion to z , at the slope of $1/2$. So this takes us up to 1.

And then it starts increasing towards 1, according to this function. When you let z go to infinity, the cumulative is going to go to 1. And it has a shape of, more or less, this kind.

So now to get the density, we just take the derivative. And the density is, of course, 0 down here. Up here the derivative is just 1/2. And beyond that point we need to take the derivative of this expression.

And the derivative is going to be 1/2 times 1 over z-squared. So it's going to be a shape of this kind. And we're done.

So you see that problems involving functions of multiple random variables are no harder than problems that deal with the functional of a single random variable. The general procedure is, again, exactly the same. You first find the cumulative, and then you differentiate. The only extra difficulty will be that when you calculate the cumulative, you need to find the probability of an event that involves multiple random variables. And sometimes this could be a little harder to do.

By the way, since we dealt with this example, just a couple of questions. What do you think is going to be the expected value of the random variable Z? Let's see, the expected value of the random variable Z is going to be the integral of z times the density.

And the density is equal to 1/2 for z going from 0 to 1. And then there's another contribution from 1 to infinity. There the density is 1/(2z-squared). And we get the z, since we're dealing with expectations, dz.

So what is this integral? Well, if you look here, you're integrating 1/z, all the way to infinity. 1/z has an integral, which is the logarithm of z. And since the logarithm goes to infinity, this means that this integral is also infinite.

So the expectation of the random variable Z is actually infinite in this example. There's nothing wrong with this. Lots of random variables have infinite expectations. If the tail of the density falls kind of slowly, as the argument goes to infinity, then it may well turn out that you get an infinite integral. So that's just how things often are. Nothing strange about it.

And now, since we are still in this example, let me ask another question. Would we reason, on the average, would it be true that the expected value of Z -- remember that Z is the ratio Y/X -- could it be that the expected value of Z is this number? Or could it be that it's equal to this number? Or could it be that it's none of the above?

OK, so how many people think this is correct? Small number. How many people think this is correct? Slightly bigger, but still a small number.

And how many people think this is correct? OK, that's-- this one wins the vote. OK, let's see.

This one is not correct, just because there's no reason it should be correct. So, in general, you cannot reason on the average. The expected value of a function is not the same as the same function of the expected values. This is only true if you're dealing with linear functions of random variables. So this is not-- this turns out to not be correct.

How about this one? Well, X and Y are independent, by assumption. So $1/X$ and Y are also independent. Why is this? Independence means that one random variable does not convey any information about the other.

So Y doesn't give you any information about X . So Y doesn't give you any information about $1/X$. Or to put it differently, if two random variables are independent, functions of each one of those random variables are also independent.

If X is independent from Y , then $g(X)$ is independent of $h(Y)$. So this applies to this case. These two random variables are independent.

And since they are independent, this means that the expected value of their product is equal to the product of the expected values. So this relation actually is true. And therefore, this is not true. OK.

Now, let's move on. We have this general procedure of finding the derived distribution by going through the cumulative. Are there some cases where we can have a shortcut? Turns out that there is a special case or a special structure in which we can get directly from densities to densities using directly just a formula. And in that case, we don't have to go through the cumulative.

And this case is also interesting, because it gives us some insight about how one density changes to a different density and what affects the shape of those densities. So the case where things easy is when the transformation from one random variable to the other is a strictly monotonic one. So there's a one-to-one relation between x 's and y 's.

Here we can reason directly in terms of densities by thinking in terms of probabilities of small intervals. So let's look at the small interval on the x -axis, like this one, when X ranges from-- where capital X ranges from a small x to a small x plus delta. So this is a small interval of length delta.

Whenever X happens to fall in this interval, the random variable Y is going to fall in a corresponding interval up there. So up there we have a corresponding interval. And these two intervals, the red and the blue interval-- this is the blue interval. And that's the red interval.

These two intervals should have the same probability. They're exactly the same event. When X falls here, $g(X)$ happens to fall in there. So we can sort of say that the probability of this little interval is the same as the probability of that little interval. And we know that probabilities of little intervals have something to do with densities.

So what is the probability of this little interval? It's the density of the random variable X , at this point, times the length of the interval. How about the probability of that interval? It's going to be the density of the random variable Y times the length of that little interval.

Now, this interval has length delta. Does that mean that this interval also has length delta? Well, not necessarily.

The length of this interval has something to do with the slope of your function g . So slope is $\frac{dy}{dx}$. Is how much-- the slope tells you how big is the y interval when you take an interval x of a certain length.

So the slope is what multiplies the length of this interval to give you the length of that interval. So the length of this interval is Δx times the slope of your function. So the length of the interval is Δx times the slope of the function, approximately.

So the probability of this interval is going to be the density of Y times the length of the interval that we are considering. So this gives us a relation between the density of X , evaluated at this point, to the density of Y , evaluated at that point. The two densities are closely related.

If these x 's are very likely to occur, then this is big, which means that that density will also be big. If these x 's are very likely to occur, then those y 's are also very likely to occur. But there's also another factor that comes in. And that's the slope of the function at this particular point.

So we have this relation between the two densities. Now, in interpreting this equation, you need to make sure what's the relation between the two variables. I have both little x 's and little y 's.

Well, this formula is true for an (x,y) pair, that they're related according to this particular function. So if I fix an x and consider the corresponding y , then the densities at those x 's and corresponding y 's will be related by that formula. Now, in the end, you want to come up with a formula that just gives you the density of Y as a function of y . And that means that you need to eliminate x from the picture.

So let's see how that would go in an example. So suppose that we're dealing with the function y equal to x cubed, in which case our function, $g(x)$, is the function x cubed. And if x cubed is equal to a little y , If we have a pair of x 's and y 's that are related this way, then this means that x is going to be the cubic root of y .

So this is the formula that takes us back from y 's to x 's. This is the direct function from x , how to construct y . This is essentially the inverse function that tells us, from a given y what is the corresponding x . Now, if we write this formula, it tells us that the density at the particular x is going to be the density at the corresponding y times the slope of the function at the particular x that we are considering. The slope of the function is $3x^2$.

Now, we want to end up with a formula for the density of Y . So I'm going to take this factor, send it to the other side. But since I want it to be a function of y , I want to eliminate the x 's. And I'm going to eliminate the x 's using this correspondence here.

So I'm going to get the density of X evaluated at y to the $1/3$. And then this factor in the denominator, it's $1/(3y)^{2/3}$. So we end up finally with the formula for the density of the random variable Y .

And this is the same answer that you would get if you go through this exercise using the cumulative distribution function method. You end up getting the same answer. But here we sort of get it directly.

Just to get a little more insight as to why the slope comes in-- suppose that we have a function like this one. So the function is sort of flat, then moves quickly, and then becomes flat again. What should be -- and suppose that X has some kind of reasonable density, some kind of flat density.

Suppose that X is a pretty uniform random variable. What's going to happen to the random variable Y? What kind of distribution should it have? What are the typical values of the random variable Y?

Either x falls here, and y is a very small number, or-- let's take that number here to be -- let's say 2 -- or x falls in this range, and y takes a value close to 2. And there's a small chance that x's will be somewhere in the middle, in which case y takes intermediate values. So what kind of shape do you expect for the distribution of Y?

There's going to be a fair amount of probability that Y takes values close to 0. There's a small probability that Y takes intermediate values. That corresponds to the case where x falls in here.

That's not a lot of probability. So the probability that Y takes values between 0 and 2, that's kind of small. But then there's a lot of x's that produces y's that are close to 2. So there's a significant probability that Y would take values that are close to 2.

So you-- the density of Y would have a shape of this kind. By looking at this picture, you can tell that it's most likely that either x will fall here or x will fall there. So the g(x) is most likely to be close to 0 or to be close to 2.

So since y is most likely to be close to 0 or close to most of the probability of y is here. And there's a small probability of being in between. Notice that the y's that get a lot of probability are those y's associated with flats regions off your g function. When the g function is flat, that gives you big densities for Y.

So the density of Y is inversely proportional to the slope of the function. And that's what you get from here. The density of Y is-- send that term to the other side-- is inversely proportional to the slope of the function that you're dealing with.

OK, so this formula works nicely for the case where the function is one-to-one. So we can have a unique association between x's and y's and through an inverse function, from y's to x's. It works for the monotonically increasing case. It also works for the monotonically decreasing case. In the monotonically decreasing case, the only change that you need to do is to take the absolute value of the slope, instead of the slope itself.

OK, now, here's another example or a special case. Let's talk about the most interesting case that involves a function of two random variables. And this is the case where we have two

independent, random variables, and we want to find the distribution of the sum of the two. We're really interested in the continuous case. But as a warm-up, it's useful to look at the discrete case first of discrete random variables.

Let's say we want to find the probability that the sum of X and Y is equal to a particular number. And to illustrate this, let's take that number to be equal to 3. What's the probability that the sum of the two random variables is equal to 3?

To find the probability that the sum is equal to 3, you consider all possible ways that you can get the sum of 3. And the different ways are the points in this picture. And they correspond to a line that goes this way. So the probability that the sum is equal to a certain number is the probability that -- is the sum of the probabilities of all of those points.

What is a typical point in this picture? In a typical point, the random variable X takes a certain value. And Y takes the value that's needed so that the sum is equal to w . Any combination of an x with a $w - x$, any such combination gives you a sum of w .

So the probability that the sum is w is the sum over all possible x 's. That's over all these points of the probability that we get a certain x . Let's say x equals 2 times the corresponding probability that random variable Y takes the value 1.

And why am I multiplying probabilities here? That's where we use the assumption that the two random variables are independent. So the probability that X takes a certain value and Y takes the complementary value, that probability is the product of two probabilities because of independence.

And when we write that into our usual PMF notation, it's a formula of this kind. So this formula is called the convolution formula. It's an operation that takes one PMF and another PMF-- p we're given the PMF's of X and Y -- and produces a new PMF.

So think of this formula as giving you a transformation. You take two PMF's, you do something with them, and you obtain a new PMF. This procedure, what this formula does is -- nicely illustrated sort of by mechanically. So let me show you a picture here and illustrate how the mechanics go, in general.

So you don't have these slides, but let's just reason through it. So suppose that you are given the PMF of X , and it has this shape. You're given the PMF of Y . It has this shape. And somehow we are going to do this calculation.

Now, we need to do this calculation for every value of W , in order to get the PMF of W . Let's start by doing the calculation just for one case. Suppose the W is equal to 0, in which case we need to find the sum of $P_x(x)$ and $P_y(-x)$.

How do you do this calculation graphically? It involves the PMF of X . But it involves the PMF of Y , with the argument reversed. So how do we plot this?

Well, in order to reverse the argument, what you need is to take this PMF and flip it. So that's where it's handy to have a pair of scissors with you. So you cut this down. And so now you take the PMF of the random variable Y and just flip it.

So what you see here is this function where the argument is being reversed. And then what do we do? We cross-multiply the two plots. Any entry here gets multiplied with the corresponding entry there. And we consider all those products and add them up.

In this particular case, the flipped PMF doesn't have any overlap with the PMF of X. So we're going to get an answer that's equal to 0. So for w's equal to 0, the P_w is going to be equal to 0, in this particular plot.

Now if we have a different value of w -- oops. If we have a different value of the argument w, then we have here the PMF of Y that's flipped and shifted by an amount of w. So the correct picture of what you do is to take this and displace it by a certain amount of w.

So here, how much did I shift it? I shifted it until one falls just below 4. So I have shifted by a total amount of 5. So 0 falls under 5, whereas 0 initially was under 0. So I'm shifting it by 5 units.

And I'm now going to cross-multiply and add. Does this give us the correct-- does it do the correct thing? Yes, because a typical term will be the probability that this random variable is 3 times the probability that this random variable is 2. That's a particular way that you can get a sum of 5.

If you see here, the way that things are aligned, it gives you all the different ways that you can get the sum of 5. You can get the sum of 5 by having 1 + 4, or 2 + 3, or 3 + 2, or 4 + 1. You need to add the probabilities of all those combinations.

So you take this times that. That's one product term. Then this times 0, this times that. And so 1-- you cross-- you find all the products of the corresponding terms, and you add them together.

So it's a kind of handy mechanical procedure for doing this calculation, especially when the PMF's are given to you in terms of a picture. So the summary of these mechanics are just what we did, is that you put the PMF's on top of each other.

You take the PMF of Y. You flip it. And for any particular w that you're interested in, you take this flipped PMF and shift it by an amount of w. Given this particular shift for a particular value of w, you cross-multiply terms and then accumulate them or add them together.

What would you expect to happen in the continuous case? Well, the story is familiar. In the continuous case, pretty much, almost always things work out the same way, except that we replace PMF's by PDF's. And we replace sums by integrals.

So there shouldn't be any surprise here that you get a formula of this kind. The density of W can be obtained from the density of X and the density of Y by calculating this integral. Essentially,

what this integral does is it fits a particular w of interest. We're interested in the probability that the random variable, capital W , takes a value equal to little w or values close to it.

So this corresponds to the event, which is this particular line on the two-dimensional space. So we need to find the sort of odd probabilities along that line. But since the setting is continuous, we will not add probabilities. We're going to integrate. And for any typical point in this picture, the probability of obtaining an outcome in this neighborhood is the-- has something to do with the density of that particular x and the density of the particular y that would compliment x , in order to form a sum of w .

So this integral that we have here is really an integral over this particular line. OK, so I'm going to skip the formal derivation of this result. There's a couple of derivations in the text. And the one which is outlined here is yet a third derivation.

But the easiest way to make sense of this formula is to consider what happens in the discrete case. So for the rest of the lecture we're going to consider a few extra, more miscellaneous topics, a few remarks, and a few more definitions. So let's change-- flip a page and consider the next mini topic.

There's not going to be anything deep here, but just something that's worth being familiar with. If you have two independent, normal random variables with certain parameters, the question is, what does the joined PDF look like? So if they're independent, by definition the joint PDF is the product of the individual PDF's.

And the PDF's each one of them involves an exponential of something. The product of two exponentials is the exponential of the sum. So you just add the exponents.

So this is the formula for the joint PDF. Now, you look at that formula and you ask, what does it look like? OK, you can understand it, a function of two variables by thinking about the contours of this function.

Look at the points at which the function takes a constant value. Where is it? When is it constant? What's the shape of the set of points where this is a constant? So consider all x 's and y 's for which this expression here is a constant, that this expression here is a constant.

What kind of shape is this? This is an ellipse. And it's an ellipse that's centered at-- it's centered at μ_x , μ_y . These are the means of the two random variables.

If those sigmas were equal, that ellipse would be actually a circle. And you would get contours of this kind. But if, on the other hand, the sigmas are different, you're going to get an ellipse that has contours of this kind.

So if my contours are of this kind, that corresponds to what? Sigma x being bigger than sigma y or vice versa. OK, contours of this kind basically tell you that X is more likely to be spread out than Y . So the range of possible x 's is bigger.

And X out here is as likely as a Y up there. So big X 's have roughly the same probability as certain smaller y 's. So in a picture of this kind, the variance of X is going to be bigger than the variance of Y .

So depending on how these variances compare with each other, that's going to determine the shape of the ellipse. If the variance of Y we're bigger, then your ellipse would be the other way. It would be elongated in the other dimension. Just visualize it a little more.

Let me throw at you a particular picture. This is one-- this is a picture of one special case. Here, I think, the variances are equal. That's the kind of shape that you get. It looks like a two-dimensional bell.

So remember, for a normal random variables, for a single random variable you get a PDF that's bell shaped. That's just a bell-shaped curve. In the two-dimensional case, we get the joint PDF, which is bell shaped again. And now it looks more like a real bell, the way it would be laid out in ordinary space.

And if you look at the contours of this function, the places where the function is equal, the typical contour would have this shape here. And it would be an ellipse. And in this case, actually, it will be more like a circle.

So these would be the different contours for different-- so the contours are places where the joint PDF is a constant. When you change the value of that constant, you get the different contours. And the PDF is, of course, centered around the mean of the two random variables. So in this particular case, since the bell is centered around the $(0, 0)$ vector, this is a plot of a bivariate normal with 0 means.

OK, there's-- bivariate normals are also interesting when your bell is oriented differently in space. We talked about ellipses that are this way, ellipses that are this way. You could imagine also bells that you take them, you squash them somehow, so that they become narrow in one dimension and then maybe rotate them.

So if you had-- we're not going to go into this subject, but if you had a joint pdf whose contours were like this, what would that correspond to? Would your x 's and y 's be independent? No.

This would indicate that there's a relation between the x 's and the y 's. That is, when you have bigger x 's, you would expect to also get bigger y 's. So it would be a case of dependent normals. And we're coming back to this point in a second.

Before we get to that point in a second that has to do with the dependencies between the random variables, let's just do another digression. If we have our two normals that are independent, as we discussed here, we can go and apply the formula, the convolution formula that we were just discussing. Suppose you want to find the distribution of the sum of these two independent normals.

How do you do this? There is a closed-form formula for the density of the sum, which is this one. We do have formulas for the density of X and the density of Y , because both of them are normal, random variables.

So you need to calculate this particular integral here. It's an integral with respect to x . And you have to calculate this integral for any given value of w .

So this is an exercise in integration, which is not very difficult. And it turns out that after you do everything, you end up with an answer that has this form. And you look at that, and you suddenly recognize, oh, this is normal. And conclusion from this exercise, once it's done, is that the sum of two independent normal random variables is also normal.

Now, the mean of W is, of course, going to be equal to the sum of the means of X and Y . In this case, in this formula I took the means to be 0. So the mean of W is also going to be 0. In the more general case, the mean of W is going to be just the sum of the two means.

The variance of W is always the sum of the variances of X and Y , since we have independent random variables. So there's no surprise here. The main surprise in this calculation is this fact here, that the sum of independent normal random variables is normal. I had mentioned this fact in a previous lecture.

Here what we're doing is to basically outline the argument that justifies this particular fact. It's an exercise in integration, where you realize that when you convolve two normal curves, you also get back a normal one once more. So now, let's return to the comment I was making here, that if you have a contour plot that has, let's say, a shape of this kind, this indicates some kind of dependence between your two random variables.

So instead of a contour plot, let me throw in here a scattered diagram. What does this scattered diagram correspond to?

Suppose you have a discrete distribution, and each one of the points in this diagram has positive probability. When you look at this diagram, what would you say? I would say that when y is big then x also tends to be larger.

So bigger x 's are sort of associated with bigger y 's in some average, statistical sense. Whereas, if you have a picture of this kind, it tells you in association that the positive y 's tend to be associated with negative x 's most of the time. Negative y 's tend to be associated mostly with positive x 's.

So here there's a relation that when one variable is large, the other one is also expected to be large. Here there's a relation of the opposite kind. How can we capture this relation between two random variables?

The way we capture it is by defining this concept called the covariance, that looks at the relation of was X bigger than usual? That's the question, whether this is positive. And how does this relate to the answer-- to the question, was Y bigger than usual?

We're asking-- by calculating this quantity, we're sort of asking the question, is there a systematic relation between having a big X with having a big Y? OK , to understand more precisely what this does, let's suppose that the random variable has 0 means, So that we get rid of this-- get rid of some clutter. So the covariance is defined just as this product.

What does this do? If positive x's tends to go together with positive y's, and negative x's tend to go together with negative y's, this product will always be positive. And the covariance will end up being positive. In particular, if you sit down with a scattered diagram and you do the calculations, you'll find that the covariance of X and Y in this diagram would be positive, because here, most of the time, X times Y is positive. There's going to be a few negative terms, but there are fewer than the positive ones.

So this is a case of a positive covariance. It indicates a positive relation between the two random variables. When one is big, the other also tends to be big.

This is the opposite situation. Here, when one variable-- here, most of the action happens in this quadrant and that quadrant, which means that X times Y, most of the time, is negative. You get a few positive contributions, but there are few. When you add things up, the negative terms dominate. And in this case we have covariance of X and Y being negative.

So a positive covariance indicates a sort of systematic relation, that there's a positive association between the two random variables. When one is large, the other also tends to be large. Negative covariance is sort of the opposite. When one tends to be large, the other variable tends to be small.

OK, so what else is there to say about the covariance? One observation to make is the following. What's the covariance of X with X itself?

If you plug in X here, you see that what we have is expected value of X minus expected of X squared. And that's just the definition of the variance of a random variable. So that's one fact to keep in mind.

We had a shortcut formula for calculating variances. There's a similar shortcut formula for calculating covariances. In particular, we can calculate covariances in this particular way. That's just the convenient way of doing it whenever you need to calculate it.

And finally, covariances are very useful when you want to calculate the variance of a sum of random variables. We know that if two random variables are independent, the variance of the sum is the sum of the variances. When the random variables are dependent, this is no longer true, and we need to supplement the formula a little bit.

And there's a typo on the slides that you have in your hands. That term of 2 shouldn't be there. And let's see where that formula comes from.

Let's suppose that our random variables are independent of -- not independent -- our random variables have 0 means. And we want to calculate the variance. So the variance is going to be

expected value of (X_1 plus X_n) squared. What you do is you expand the square. And you get the expected value of the sum of the X_i squared.

And then you get all the cross terms. OK. And so now, here, let's assume for simplicity that we have 0 means. The expected value of this is the sum of the expected values of the X squared terms. And that gives us the variance.

And then we have all the possible cross terms. And each one of the possible cross terms is the expected value of X_i times X_j . This is just the covariance.

So if you can calculate all the variances and the covariances, then you're able to calculate also the variance of a sum of random variables. Now, if two random variables are independent, then you look at this expression. Because of independence, expected value of the product is going to be the product of the expected values. And the expected value of just this term is always equal to 0.

You're expected deviation from the mean is just 0. So the covariance will turn out to be 0. So independent random variables lead to 0 covariances, although the opposite fact is not necessarily true. So covariances give you some indication of the relation between two random variables.

Something that's not so convenient conceptually about covariances is that it has the wrong units. That's the same comment that we had made regarding variances. And with variances we got out of that issue by considering the standard deviation, which has the correct units.

So with the same reasoning, we want to have a concept that captures the relation between two random variables and, in some sense, that doesn't have to do with the units that we're dealing. We want to have a dimensionless quantity. That tells us how strongly two random variables are related to each other.

So instead of considering the covariance of just X with Y , we take our random variables and standardize them by dividing them by their individual standard deviations and take the expectation of this. So what we end up doing is the covariance of X and Y , which has units that are the units of X times the units of Y . But divide with a standard deviation, so that we get a quantity that doesn't have units.

This quantity, we call it the correlation coefficient. And it's a very useful quantity, a very useful measure of the strength of association between two random variables. It's very informative, because it falls always between -1 and +1. This is an algebraic exercise that you're going to see in recitation.

And the way that you interpret it is as follows. If the two random variables are independent, the covariance is going to be 0. The correlation coefficient is going to be 0. So 0 correlation coefficient basically indicates a lack of a systematic relation between the two random variables.

On the other hand, when rho is large, either close to 1 or close to -1, this is an indication of a strong association between the two random variables. And the extreme case is when rho takes an extreme value.

When rho has a magnitude equal to 1, it's as big as it can be. In that case, the two random variables are very strongly related. How strongly? Well, if you know one random variable, if you know the value of y, you can recover the value of x and conversely.

So the case of a complete correlation is the case where one random variable is a linear function of the other random variable. In terms of a scatter plot, this would mean that there's a certain line and that the only possible (x,y) pairs that can happen would lie on that line. So if all the possible (x,y) pairs lie on this line, then you have this relation, and the correlation coefficient is equal to 1. A case where the correlation coefficient is close to 1 would be a scatter plot like this, where the x's and y's are quite strongly aligned with each other, maybe not exactly, but fairly strongly. All right, so you're going to hear a little more about correlation coefficients and covariances in recitation tomorrow.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013

Transcript – Recitation: The Difference of Two Independent Exponential Random Variables

In this problem, Romeo and Juliet are to meet up for a date, where Romeo arrives at time x and Juliet at time y , where x and y are independent exponential random variables, with parameters λ . And we're interested in knowing the difference between the two times of arrivals, we'll call it z , written as x minus y . And we'll like to know what the distribution of z is, expressed by the probability density function, f of z .

Now, we'll do so by using the so-called convolution formula that we learn in the lecture. Recall that if we have a random variable w that is the sum of two independent random variables, x plus y , now, if that's the case, we can write the probability [INAUDIBLE] function, f_w , [INAUDIBLE] as the following integration-- negative infinity to infinity f_x little x times f_y w minus x , integrated over x .

And to use this expression to calculate f of z , we need to do a bit more work. Notice w is expressed as a sum of two random variables, whereas z is expressed as the subtraction of y from x . But that's fairly easy to fix. Now, we can write z . Instead of a subtraction, write it as addition of x plus negative y .

So in the expression of the convolution formula, we'll simply replace y by negative y , as it will show on the next slide. Using the convolution formula, we can write f of z little z as the integration of f of x little x and f of negative y z minus x dx . Now, we will use the fact that f of negative y , evaluated z minus x , is simply equal to f of y evaluated at x minus z .

To see why this is true, let's consider, let's say, a discrete random variable, y . And now, the probability that negative y is equal to negative 1 is simply the same as probability that y is equal to 1. And the same is true for probability density functions. With this fact in mind, we can further write equality as the integration of x times f of y x minus z dx .

We're now ready to compute. We'll first look at the case where z is less than 0. On the right, I'm writing out the distribution of an exponential random variable with a parameter λ . In this case, using the integration above, we could write it as 0 to infinity, $\lambda e^{-\lambda z}$ dx .

Now, the reason we chose a region to integrate from 0 to positive infinity is because anywhere else, as we can verify from the expression of f_x right here, that the product of f_x times f_y here is 0. Follow this through.

We'll pull out the constant. $\lambda e^{-\lambda z}$, the integral from 0 to infinity, $\lambda e^{-\lambda z}$ to the negative 2 λx dx . This will give us $\lambda e^{-\lambda z}$ minus 1/2 $e^{-\lambda z}$ to the negative 2 λx infinity minus this expression value at 0.

And this will give us lambda over 2 e to the lambda z. So now, we have an expression for f of z evaluated at little z when little z is less than 0. Now that have the distribution of f of z when z is less than 0, we'd like to know what happens when z is greater or equal to 0. In principle, we can go through the same procedure of integration and calculate that value.

But it turns out, there's something much simpler. z is the difference between x and y, at negative z, simply the difference between y and x. Now, x and y are independent and identically distributed. And therefore, x minus y has the same distribution as y minus x.

So that tells us z and negative z have the same distribution. What that means is, is the distribution of z now must be symmetric around 0. In other words, if we know that the shape of f of z below 0 is something like that, then the shape of it above 0 must be symmetric. So here's the origin.

For example, if we were to evaluate f of z at 1, well, this will be equal to the value of f of z at negative 1. So this will equal to f of z at negative 1. Well, with this information in mind, we know that in general, f of z little z is equal to f of z negative little z. So what this allows us to do is to get all the information for z less than 0 and generalize it to the case where z is greater or equal to 0.

In particular, by the symmetry here, we can write, for the case z greater or equal to 0, as lambda over 2 e to the negative lambda z. So the negative sign comes from the fact that the distribution of f of z is symmetric around 0. And simply, we can go back to the expression here to get the value. And all in all, this implies that f of z little z is equal to lambda over 2 e to the negative lambda absolute value of z.

This completes our problem.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013

Transcript – Tutorial: The Sum of Discrete and Continuous Random Variables

In this video, we're going to do an example in which we derive the probability density function of the sum of two random variables.

The problem tells us the following. We're given that X and Y are independent random variables. X is a discrete random variable with PMF P_x . Y is continuous with PDF F_y . And we'd like to compute the PDF of Z which is equal to X plus Y . We're going to use the standard approach here-- compute the CDF of Z and then take the derivative to get the PDF.

So in this case, the CDF, which is F_z , by definition is the random variable Z being less than little z . But Z is just X plus Y . So now, we'd actually like to, instead of having to deal with two random variables, X and Y , we'd like to deal with one at a time.

And the total probability theorem allows us to do this by conditioning on one of the two random variables. Conditioning on Y here is a bit tricky, because Y is continuous, and you have to be careful with your definitions. So conditioning on X seems like the way to go. So let's do that.

This is just the probability that X equals little x , which is exactly equal to the PMF of X evaluated at x . Now we're given we're fixing X equal to little x . So we can actually replace every instance of the random variable with little x . And now I'm going to just rearrange this so that it looks a little nicer. So I'm going to have Y on the left and say Y is less than z minus x , where z minus x is just a constant.

Now, remember that X and Y are independent. So telling us something about X shouldn't change our beliefs about Y . So in this case, we can actually drop the conditioning. And this is exactly the CDF of Y evaluated at z minus x . So now we've simplified as far as we could. So let's take the derivative and see where that takes us.

So the PDF of Z is, by definition, the derivative of the CDF, which we just computed here. This is sum over x $F_y(z - x) P_x$. What next? Interchange the derivative and the summation.

And a note of caution here. So if x took on a finite number of values, you'd have a finite number of terms here. And this would be completely valid. You can just do this.

But if x took on, for example, a countably infinite number of values-- a geometric random variable, for example-- this would actually require some formal justification. But I'm not going to get into that.

So here, the derivative with respect to z -- this is actually z -- is you use chain rule here. P_x doesn't matter, because it's not a function of z . So we have F_y evaluated at z minus x according to the

chain rule, and then the derivative of the inner quantity, z minus x , which is just 1. So we don't need to put anything there. And we get P_x of x .

So there we go. We've derived the PDF of Z . Notice that this looks quite similar to the convolution formula when you assume that both X and Y are either continuous or discrete. And so that tells us that this looks right.

So in summary, we've basically computed the PDF of X plus Y where X is discrete and Y is continuous. And we've used the standard two-step approach-- compute the CDF and then take the derivative to get the PDF.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

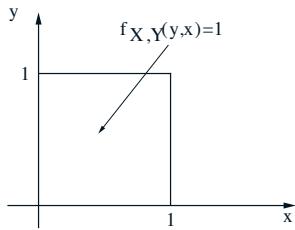
LECTURE 11

Derived distributions; convolution; covariance and correlation

- Readings:

Finish Section 4.1;
Section 4.2

Example



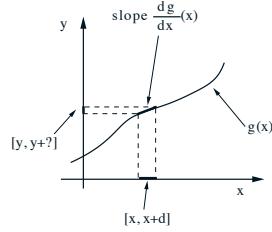
Find the PDF of $Z = g(X, Y) = Y/X$

$$F_Z(z) = \begin{cases} 0 & z \leq 1 \\ 1 & z \geq 1 \end{cases}$$

$$F_Z(z) = \begin{cases} 0 & z \leq 1 \\ 1 & z \geq 1 \end{cases}$$

A general formula

- Let $Y = g(X)$
 g strictly monotonic.



- Event $x \leq X \leq x + \delta$ is the same as
 $g(x) \leq Y \leq g(x + \delta)$
or (approximately)
 $g(x) \leq Y \leq g(x) + \delta |(dg/dx)(x)|$

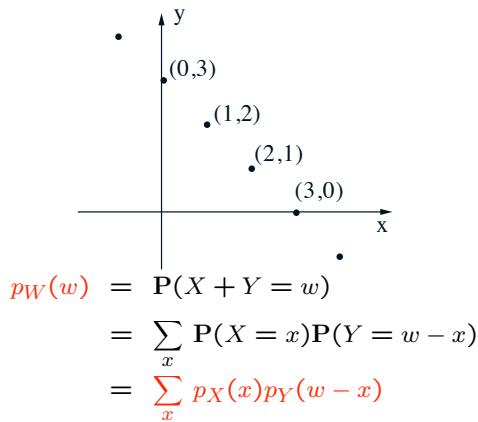
- Hence,

$$\delta f_X(x) = \delta f_Y(y) \left| \frac{dg}{dx}(x) \right|$$

where $y = g(x)$

The distribution of $X + Y$

- $W = X + Y$; X, Y independent

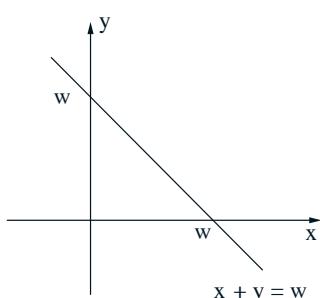


- Mechanics:

- Put the pmf's on top of each other
- Flip the pmf of Y
- Shift the flipped pmf by w
(to the right if $w > 0$)
- Cross-multiply and add

The continuous case

- $W = X + Y$; X, Y independent



- $f_{W|X}(w | x) = f_Y(w - x)$
- $f_{W,X}(w, x) = f_X(x)f_{W|X}(w | x)$
 $= f_X(x)f_Y(w - x)$
- $f_W(w) = \int_{-\infty}^{\infty} f_X(x)f_Y(w - x) dx$

Two independent normal r.v.s

- $X \sim N(\mu_x, \sigma_x^2)$, $Y \sim N(\mu_y, \sigma_y^2)$,
independent

$$f_{X,Y}(x,y) = f_X(x)f_Y(y)$$

$$= \frac{1}{2\pi\sigma_x\sigma_y} \exp \left\{ -\frac{(x-\mu_x)^2}{2\sigma_x^2} - \frac{(y-\mu_y)^2}{2\sigma_y^2} \right\}$$

- PDF is constant on the ellipse where

$$\frac{(x-\mu_x)^2}{2\sigma_x^2} + \frac{(y-\mu_y)^2}{2\sigma_y^2}$$

is constant

- Ellipse is a circle when $\sigma_x = \sigma_y$

The sum of independent normal r.v.'s

- $X \sim N(0, \sigma_x^2)$, $Y \sim N(0, \sigma_y^2)$,
independent

- Let $W = X + Y$

$$f_W(w) = \int_{-\infty}^{\infty} f_X(x)f_Y(w-x) dx$$

$$= \frac{1}{2\pi\sigma_x\sigma_y} \int_{-\infty}^{\infty} e^{-x^2/2\sigma_x^2} e^{-(w-x)^2/2\sigma_y^2} dx$$

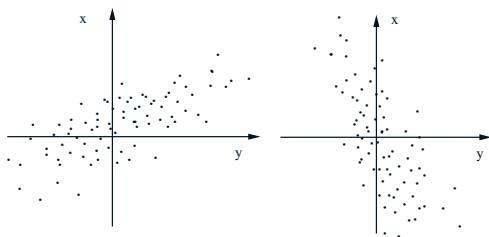
$$(\text{algebra}) = ce^{-\gamma w^2}$$

- Conclusion: W is normal

- mean=0, variance= $\sigma_x^2 + \sigma_y^2$
- same argument for nonzero mean case

Covariance

- $\text{cov}(X, Y) = E[(X - E[X]) \cdot (Y - E[Y])]$
- Zero-mean case: $\text{cov}(X, Y) = E[XY]$



- $\text{cov}(X, Y) = E[XY] - E[X]E[Y]$

$$\text{var} \left(\sum_{i=1}^n X_i \right) = \sum_{i=1}^n \text{var}(X_i) + 2 \sum_{(i,j):i \neq j} \text{cov}(X_i, X_j)$$

- independent $\Rightarrow \text{cov}(X, Y) = 0$
(converse is not true)

Correlation coefficient

- Dimensionless version of covariance:

$$\rho = E \left[\frac{(X - E[X])}{\sigma_X} \cdot \frac{(Y - E[Y])}{\sigma_Y} \right]$$

$$= \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

- $-1 \leq \rho \leq 1$

- $|\rho| = 1 \Leftrightarrow (X - E[X]) = c(Y - E[Y])$
(linearly related)

- Independent $\Rightarrow \rho = 0$
(converse is not true)

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Recitation 12
October 19, 2010

1. Show $\rho(aX + b, Y) = \rho(X, Y)$.
2. Romeo and Juliet have a date at a given time, and each, independently, will be late by amounts of time, X and Y , respectively, that are exponentially distributed with parameter λ .
 - (a) Find the PDF of $Z = X - Y$ by first finding the CDF and then differentiating.
 - (b) Find the PDF of Z by using the total probability theorem.
3. Problem 4.16, page 248 in text.

Let X and Y be independent standard normal random variables. The pair (X, Y) can be described in polar coordinates in terms of random variables $R \geq 0$ and $\Theta \in [0, 2\pi]$, so that

$$X = R\cos\Theta, \quad Y = R\sin\Theta.$$

Show that R and Θ are independent (i.e. show $f_{R,\Theta}(r, \theta) = f_R(r)f_\Theta(\theta)$).

- (a) Find $f_R(r)$.
 - (b) Find $f_\Theta(\theta)$.
 - (c) Find $f_{R,\Theta}(r, \theta)$.
4. Problem 4.20, page 250 in text. **Schwarz inequality**.
Show that for any random variables X and Y , we have

$$(\mathbf{E}[XY])^2 \leq \mathbf{E}[X^2]\mathbf{E}[Y^2].$$

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Recitation 12 Solutions

October 19, 2010

1.

$$\begin{aligned}
 \rho(aX + b, Y) &= \frac{\text{cov}(aX + b, Y)}{\sqrt{\text{var}(aX + b)\text{var}(Y)}} \\
 &= \frac{\mathbf{E}[(aX + b - \mathbf{E}[aX + b])(Y - \mathbf{E}[Y])]}{\sqrt{a^2\text{var}(X)\text{var}(Y)}} \\
 &= \frac{\mathbf{E}[(aX + b - a\mathbf{E}[X] - b)(Y - \mathbf{E}[Y])]}{a\sqrt{\text{var}(X)\text{var}(Y)}} \\
 &= \frac{a\mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])]}{a\sqrt{\text{var}(X)\text{var}(Y)}} \\
 &= \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)\text{var}(Y)}} \\
 &= \rho(X, Y)
 \end{aligned}$$

As an example where this property of the correlation coefficient is relevant, consider the homework and exam scores of students in a class. We expect the homework and exam scores to be positively correlated and thus have a positive correlation coefficient. Note that, in this example, the above property will mean that the correlation coefficient will not change whether the exam is out of 105 points, 10 points, or any other number of points.

2.

(a) When $z \geq 0$:

$$\begin{aligned}
 F_Z(z) = \mathbf{P}(X - Y \leq z) &= \mathbf{P}(X \leq Y + z) \\
 &= \int_0^\infty \int_0^{y+z} f_{X,Y}(x, y') dx dy \\
 &= \int_0^\infty \lambda e^{-\lambda y} \int_0^{y+z} \lambda e^{-\lambda x} dx dy \\
 &= \lambda^y \left(1 - e^{-\lambda(y+z)}\right) dy \\
 &= 1 + \frac{e^{-\lambda z}}{2} e^{-2\lambda y} \Big|_{y=0}^{y=\infty} \\
 &= 1 - \frac{1}{2} e^{-\lambda z} \quad z \geq 0
 \end{aligned}$$

When $z < 0$:

$$\begin{aligned}
 \mathbf{P}(X \leq Y + z) &= \int_0^\infty \int_0^{y+z} f_{X,Y}(x,y) dx dy \\
 &= \int_0^\infty \lambda e^{-\lambda x} \int_{x-z}^\infty \lambda e^{-\lambda y} dy dx \\
 &= \int_0^\infty \lambda e^{-\lambda x} e^{-\lambda(x-z)} dx \\
 &= e^{\lambda z} \int_0^\infty \lambda e^{-2\lambda x} dx \\
 &= \frac{1}{2} e^{\lambda z} \quad z \leq 0
 \end{aligned}$$

$$F_Z(z) = \begin{cases} 1 - \frac{1}{2} e^{-\lambda z} & z \geq 0 \\ \frac{1}{2} e^{\lambda z} & z < 0 \end{cases}$$

$$f_Z(z) = \frac{d}{dz} F_Z(z) = \begin{cases} \frac{\lambda}{2} e^{-\lambda z} & z \geq 0 \\ \frac{\lambda}{2} e^{\lambda z} & z < 0 \end{cases}$$

Hence,

$$f_Z(z) = \frac{\lambda}{2} e^{\lambda|z|}$$

(b) Solving using the total probability theorem, we have:

$$\begin{aligned}
 f_Z(z) &= \int_{-\infty}^\infty f_X(x) f_{Z|X}(z|x) dx \\
 &= \int_{-\infty}^\infty f_X(x) f_{Y|X}(x-z|x) dx \\
 &= \int_{-\infty}^\infty f_X(x) f_Y(x-z) dx
 \end{aligned}$$

First when $z < 0$, we have:

$$\begin{aligned}
 \int_{-\infty}^\infty f_X(x) f_Y(x-z) dx &= \int_0^\infty \lambda e^{-\lambda x} \lambda e^{-\lambda(x-z)} dx \\
 &= \lambda e^{\lambda z} \int_0^\infty \lambda e^{-2\lambda x} dx \\
 &= \frac{\lambda}{2} e^{\lambda z}
 \end{aligned}$$

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
 (Fall 2010)

Then, when $z \geq 0$ we have:

$$\begin{aligned} \int_{-\infty}^{\infty} f_X(x)f_Y(x-z)dx &= \int_z^{\infty} \lambda e^{-\lambda x} \lambda e^{-\lambda(x-z)} dx \\ &= \lambda e^{\lambda z} \int_z^{\infty} \lambda e^{-2\lambda x} dx \\ &= \frac{\lambda}{2} e^{\lambda z} e^{-2\lambda z} \\ &= \frac{\lambda}{2} e^{-\lambda z} \\ f_Z(z) &= \frac{\lambda}{2} e^{-\lambda|z|} \quad \forall z \end{aligned}$$

3. (a) We have $X = R\cos(\Theta)$ and $Y = R\sin(\Theta)$. Recall that in polar coordinates, the differential area is $dA = dx dy = r dr d\theta$. So

$$\begin{aligned} F_R(r) = \mathbf{P}(R \leq r) &= \int_0^r \int_0^{2\pi} f_X(r'\cos\theta) f_Y(r'\sin\theta) d\theta r' dr' \\ &= \int_0^r \int_0^{2\pi} \frac{1}{2\pi} e^{-(r')^2/2} d\theta r' dr' \\ &= \int_0^r r' e^{-(r')^2/2} dr' \int_0^{2\pi} \frac{d\theta}{2\pi} \\ &= \int_0^{r^2/2} e^{-u} du \quad (u = (r')^2/2) \\ F_R(r) &= \begin{cases} 1 - e^{-r^2/2} & r \geq 0 \\ 0 & r < 0 \end{cases} \end{aligned}$$

$$\begin{aligned} f_R(r) = \frac{d}{dr} F_R(r) &= (-1/2)(2r)(-e^{-r^2/2}) \\ &= r e^{-r^2/2}, \quad r \geq 0 \end{aligned}$$

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
 (Fall 2010)

(b)

$$\begin{aligned}
 F_\Theta(\theta) = \mathbf{P}(\Theta \leq \theta) &= \int_0^\theta \int_0^\infty f_X(r\cos\theta') f_Y(r\sin\theta') r dr d\theta' \\
 &= \int_0^\theta \int_0^\infty \frac{1}{2\pi} e^{-r^2/2} r dr d\theta' \\
 &= \int_0^\infty r e^{-r^2/2} dr \int_0^\theta \frac{d\theta'}{2\pi} \\
 &= \frac{\theta}{2\pi} \int_0^\infty e^{-u} du \quad (u = r^2/2) \\
 &= \frac{\theta}{2\pi} (-e^{-u}) \Big|_0^\infty = \frac{\theta}{2\pi} \quad 0 \leq \theta \leq 2\pi
 \end{aligned}$$

$$F_\Theta(\theta) = \begin{cases} 0 & \theta < 0 \\ \frac{\theta}{2\pi} & 0 \leq \theta \leq 2\pi \\ 1 & \theta \geq 2\pi \end{cases}$$

$$f_\Theta(\theta) = \frac{d}{d\theta} F_\Theta(\theta) = \frac{1}{2\pi} \quad 0 \leq \theta \leq 2\pi$$

(c)

$$\begin{aligned}
 F_{R,\Theta}(r, \theta) = P(R \leq r, \Theta \leq \theta) &= \int_0^\theta \int_0^r \frac{1}{2\pi} r' e^{-(r')^2/2} dr' d\theta' \\
 &= \int_0^\theta \int_0^{r^2/2} \frac{1}{2\pi} e^{-u} du d\theta' \quad (u = (r')^2/2) \\
 &= \int_0^\theta \frac{1}{2\pi} \left(1 - e^{-r^2/2}\right) d\theta' \\
 &= \frac{\theta}{2\pi} \left(1 - e^{-r^2/2}\right) \quad r \geq 0, \quad \theta > 2\pi
 \end{aligned}$$

$$F_{R,\Theta}(r, \theta) = \begin{cases} \frac{\theta}{2\pi} \left(1 - e^{-r^2/2}\right) & r \geq 0, \quad 0 \leq \theta \leq 2\pi \\ 1 - e^{-r^2/2} & r \geq 0, \quad \theta > 2\pi \\ 0 & \text{otherwise} \end{cases}$$

$$f_{R,\Theta}(r, \theta) = \frac{\partial}{\partial r} \frac{\partial}{\partial \theta} F_{R,\Theta}(r, \theta) = \frac{1}{2\pi} r e^{-r^2/2} \quad r \geq 0, \quad 0 \leq \theta \leq 2\pi$$

Note: The PDF of R^2 is exponentially distributed with parameter $\lambda = 1/2$. This is a very convenient way to generate normal random variables from independent uniform and exponential random variables. We can generate an arbitrary random variable X with CDF F_X by first generating a uniform random variable and then passing the samples from the uniform distribution through the function F_X^{-1} . But since we don't have a closed-form expression for the CDF of a normal random variable, this method doesn't work. However,

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
 (Fall 2010)

we do have a closed-form expression for the exponential distribution. Therefore, we can generate an exponential distribution with parameter 1/2 and we can generate a uniform distribution in $[0, 2\pi]$, and with these two distributions we can generate standard normal distributions.

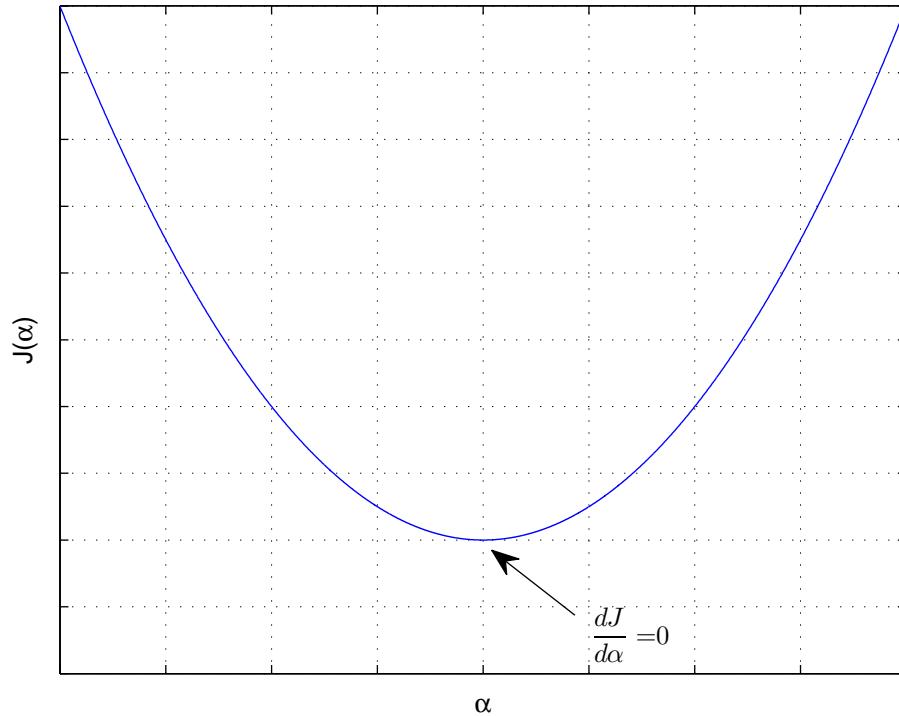
4. Problem 4.20, page 250 in text. See text for the proof.

An alternative proof is given below:

Consider the problem of picking a parameter α to minimize the expected squared difference between two random variables X and Y . Consider

$$J(\alpha) = \mathbf{E}[(X - \alpha Y)^2]$$

with $Y \neq 0$. We start with a variational calculation to find α that minimizes $J(\alpha)$. The value of α which minimizes $J(\alpha)$ is found by setting the first derivative of $J(\alpha)$ to zero (since, for $Y \neq 0$, $\frac{d^2}{d\alpha^2} J(\alpha) = 2\mathbf{E}[Y^2] > 0$).



$$\frac{d}{d\alpha} J(\alpha) = \frac{d}{d\alpha} (\mathbf{E}[X^2] - 2\alpha \mathbf{E}[XY] + \alpha^2 \mathbf{E}[Y^2]) = 0$$

$$\rightarrow \alpha = \frac{\mathbf{E}[XY]}{\mathbf{E}[Y^2]} \text{ minimizes } J(\alpha).$$

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Fall 2010)

Then

$$\begin{aligned} J\left(\frac{\mathbf{E}[XY]}{\mathbf{E}[Y^2]}\right) &= \mathbf{E}\left[\left(X - \frac{\mathbf{E}[XY]}{\mathbf{E}[Y^2]}Y\right)^2\right] \geq 0 \\ &= \mathbf{E}[X^2] - 2\frac{(\mathbf{E}[XY])^2}{\mathbf{E}[Y^2]} + \frac{(\mathbf{E}[XY])^2\mathbf{E}[Y^2]}{(\mathbf{E}[Y^2])^2} \\ &= \mathbf{E}[X^2] - \frac{(\mathbf{E}[XY])^2}{\mathbf{E}[Y^2]} \geq 0 \end{aligned}$$

Rearranging this expression gives the Schwarz inequality for expected values:

$$\mathbf{E}[X^2]\mathbf{E}[Y^2] \geq (\mathbf{E}[XY])^2$$

Note that in the above derivation, we assumed $Y \neq 0$ so that $\mathbf{E}[Y^2] > 0$. If we assume $Y = 0$ then the Schwarz inequality will hold with equality since then $\mathbf{E}[XY] = 0$ and $\mathbf{E}[Y^2] = 0$.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Fall 2010)

Tutorial 6

October 21/22, 2010

1. Let X be a discrete random variable with PMF p_X and let Y be a continuous random variable, independent from X , with PDF f_Y . Derive a formula for the PDF of the random variable $X+Y$.
 2. The random variables X and Y are described by a joint PDF which is constant within the unit area quadrilateral with vertices $(0,0)$, $(0,1)$, $(1,2)$, and $(1,1)$. Use the law of total variance to find the variance of $X + Y$.
 3. (a) You roll a fair six-sided die, and then you flip a fair coin the number of times shown by the die. Find the expected value and the variance of the number of heads obtained.
(b) Repeat part (a) for the case where you roll two dice, instead of one.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Tutorial 6: Solutions

1. Let $Z = X + Y$. Using the 2 step CDF method,

$$\begin{aligned} F_Z(z) &= \mathbf{P}(Z \leq z) \\ &= \mathbf{P}(X + Y \leq z) \end{aligned}$$

Using the Total Probability Theorem, we have

$$\begin{aligned} F_Z(z) &= \sum_x p_X(x)p(x + Y \leq z) \\ &= \sum_x p_X(x)p(Y \leq z - x) \\ &= \sum_x p_X(x)F_Y(z - x) \end{aligned}$$

Differentiating both sides with respect to z , we obtain

$$\begin{aligned} f_Z(z) &= \frac{d}{dz}F_Z(z) \\ &= \sum_x p_X(x)f_Y(z - x) \end{aligned}$$

2. We will condition on X and use the law of total variance

$$\text{var}(X + Y) = \mathbf{E}[\text{var}(X + Y|X)] + \text{var}(\mathbf{E}[X + Y|X]).$$

Given a value x of X , the random variable Y is uniformly distributed in the interval $[x, x + 1]$, and the random variable $X + Y$ is uniformly distributed in the interval $[2x, 2x + 1]$. Therefore, $\mathbf{E}[X + Y|X] = 0.5 + 2X$ and $\text{var}(X + Y|X) = 1/12$. Thus,

$$\text{var}(X + Y) = \text{var}(0.5 + 2X) + \mathbf{E}[1/12] = 4\text{var}(X) + \mathbf{E}[1/12] = \frac{5}{12}.$$

3. (a) Let X_i be independent Bernoulli random variables that are equal to 1 if the i th flip results in heads. Let N be the number of coin flips. We have $\mathbf{E}[X_i] = 1/2$, $\text{var}(X_i) = 1/4$, $\mathbf{E}[N] = 7/2$, and $\text{var}(N) = 35/12$. (The last equality is obtained from the formula for the variance of a discrete uniform random variable.) Therefore, the expected number of heads is

$$\mathbf{E}[X_i]\mathbf{E}[N] = \frac{7}{4},$$

and the variance is

$$\text{var}(X_i)\mathbf{E}[N] + (\mathbf{E}[X_i])^2\text{var}(N) = \frac{1}{4} \cdot \frac{7}{2} + \frac{1}{4} \cdot \frac{35}{12} = \frac{77}{48}.$$

- (b) The experiment in part (b) can be viewed as consisting of two independent repetitions of the experiment in part (a). Thus, both the mean and the variance are doubled and become $7/2$ and $77/24$, respectively.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013

Transcript – Tutorial: A Coin with Random Bias

Hi. In this problem, we're going to be dealing with a variation of the usual coin-flipping problem. But in this case, the bias itself of the coin is going to be random. So you could think of it as, you don't even know what the probability of heads for the coin is.

So as usual, we're still taking one coin and we're flipping it n times. But the difference here is that the bias is because it was random variable Q . And we're told that the expectation of this bias is some μ and that the variance of the bias is some σ^2 , which we're told is positive. And what we're going to be asked is find a bunch of different expectations, covariances, and variances.

And we'll see that this problem gives us some good exercise in a few concepts, a lot of iterated expectations, which, again, tells you that when you take the expectation of a conditional expectation, it's just the expectation of the inner random variable. The covariance of two random variables is just the expectation of the product minus the product of the expectations. Law of total variance is the expectation of a variance, of a conditional variance plus the variance of a conditional expectation. And the last thing, of course, we're dealing with a bunch of Bernoulli random variables, coin flips. So as a reminder, for a Bernoulli random variable, if you know what the bias is, it's some known quantity p , then the expectation of the Bernoulli is just p , and the variance of the Bernoulli is p times $1 - p$.

So let's get started. The problem tells us that we're going to define some random variables. So x_i is going to be a Bernoulli random variable for the i coin flip.

So x_i is going to be 1 if the i coin flip was heads and 0 if it was tails. And one very important thing that the problem states is that conditional on Q , the random bias, so if we know what the random bias is, then all the coin flips are independent. And that's going to be important for us when we calculate all these values.

OK, so the first thing that we need to calculate is the expectation of each of these individual Bernoulli random variables, x_i . So how do we go about calculating what this is? Well, the problem gives us a hint. It tells us to try using the law of iterated expectations. But in order to use it, you need to figure out what you need the condition on.

What this y ? What takes place in y ? And in this case, a good candidate for what you condition on would be the bias, the Q that we're unsure about. So let's try doing that and see what we get.

So we write out the law of iterated expectations with Q . So now hopefully, we can simplify it with this inter-conditional expectation is. Well, what is it really? It's saying, given what Q is, what is the expectation of this Bernoulli random variable x_i ?

Well, we know that if we knew what the bias was, then the expectation is just the bias itself. But in this case, the bias is random. But remember a conditional expectation is still a random variable.

And so in this case, this actually just simplifies into Q. So whatever the bias is, the expectation is just equal to the bias. And so that's what it tells us. And this part is easy because we're given that the expectation of q is mu.

And then the problem also defines the random variable x. X is the total number of heads within the n tosses. Or you can think of it as a sum of all these individual x_i Bernoulli random variables. And now, what can we do with this? Well we can remember that linearity of expectations allows us to split up this sum. Expectation of a sum, we could split up into a sum of expectations.

So this is actually just expectation of x_1 plus dot dot dot plus all the way to expectation of x_n . All right. And now, remember that we're flipping the same coin. We don't know what the bias is, but for all the n flips, it's the same coin. And so each of these expectations of x_i should be the same, no matter what x_i is.

And each one of them is mu. We already calculated that earlier. And there's 10 of them, so the answer would be n times mu.

So let's move on to part B. Part B now asks us to find what the covariance is between x_i and x_j . And we have to be a little bit careful here because there are two different scenarios, one where i and j are different indices, different tosses, and another where i and j are the same. So we have to consider both of these cases separately.

Let's first do the case where x and i are different. So $i \neq j$. In this case, we can just apply the formula that we talked about in the beginning. So this covariance is just equal to the expectation of $x_i x_j$ minus the expectation of x_i times expectation of x_j .

All right, so we actually know what these two are, right? Expectation of x_i is mu. Expectation of x_j is also mu. So this part is just mu squared. But we need to figure out what this expectation of $x_i x_j$ is.

Well, the expectation of $x_i x_j$, we can again use the law of iterated expectations. So let's try conditioning on cue again. And remember we said that this second part is just mu squared.

All right, well, how can we simplify this inner-conditional expectation? Well, we can use the fact that the problem tells us that, conditioned on Q, the tosses are independent. So that means that, conditioned on Q, x_i and x_j are independent.

And remember, when random variables are independent, the expectation of product, you could simplify that to be the product of the expectations. And because we're in the condition world on Q, you have to remember that it's going to be a product of two conditional expectations. So this will be expectation of x_i given Q times expectation of x_j given Q minus mu squared still.

All right, now what is this? Well the expectation of x_i given Q , we already argued earlier here that it should just be Q . And then the same thing for x_j . That should also be Q . So this is just expectation of Q squared minus μ squared.

All right, now if we look at this, what is the expectation of Q squared minus μ squared? Well, remember μ is just, we're told that μ is the expectation of Q . So what we have is the expectation of Q squared minus the quantity expectation of Q squared.

And what is that, exactly? That is just the formula or the definition of what the variance of Q should be. So this is, in fact, exactly equal to the variance of Q , which we're told is σ^2 .

All right, so what we found is that for $i \neq j$, the covariance of x_i and x_j is exactly equal to σ^2 . And remember, we're told that σ^2 is positive. So what does that tell us? That tells us that x_i and x_j , or $i \neq j$, these two random variables are correlated.

And so, because they're correlated, they can't be independent. Remember, if two intervals are independent, that means they're uncorrelated. But the converse isn't true. But if we do know that two random variables are correlated, that means that they can't be independent.

And now let's finish this by considering the second case. The second case is when i actually does equal j . And in that case, well, the covariance of x_i and x_i is just another way of writing the variance of x_i . So covariance, x_i, x_i , it's just the variance of x_i .

And what is that? That is just the expectation of x_i squared minus expectation of x_i quantity squared. And again, we know what the second term is. The second term is expectation of x_i quantity squared. Expectation of x_i we know from part A is just μ , right? So that's just second term is just μ squared.

But what is the expectation of x_i squared? Well, we can think about this a little bit more. And you can realize that x_i squared is actually exactly the same thing as just x_i .

And this is just a special case because x_i is a Bernoulli random variable. Because Bernoulli is either 0 or 1. And if it's 0 and you square it, it's still 0. And if it's 1 and you square it, it's still 1.

So squaring it doesn't really change anything. It's exactly the same thing as the original random variable. And so, because this is a Bernoulli random variable, this is exactly just the expectation of x_i .

And we said this part is just μ squared. So this is just expectation of x_i , which we said was μ . So the answer is just μ minus μ squared.

OK, so this completes part B. And the answer that we wanted was that in fact, x_i and x_j are in fact not independent. Right.

So let's write down some facts that we'll want to remember. One of them is that expectation of x_i is μ . And we also want to remember what this covariance is.

The covariance of x_i and x_j is equal to σ^2 when $i \neq j$. So we'll be using these facts again later. And the variance of x_i is equal to $\mu - \mu^2$.

So now let's move on to the last part, part C, which asks us to calculate the variance of x in two different ways. So the first way we'll do it is using the law of total variance. So the law of total variance will tell us that we can write the variance of x as a sum of two different parts. So the first is variance of x expectation of the variance of x conditioned on something plus the variance of the initial expectation of x conditioned on something. And as you might have guessed, what we're going to condition on is Q .

Let's calculate what these two things are. So let's do the two terms separately. What is the expectation of the conditional variance of x given Q ?

Well, what is-- this, we can write out x . Because x , remember, is just the sum of a bunch of these Bernoulli random variables. And now what we'll do was, well, again, use the important fact that the x 's, we're told, are conditionally independent, conditional on Q .

And because they're independent, remember the variance of a sum is not the sum of the variance. It's only the sum of the variance if the terms in the sum are independent. In this case, they are conditionally independent given Q . So we can in fact split this up and write it as the variance of x_1 given Q plus all the way to the variance of x_n given Q .

And in fact, all these are the same, right? So we just have n copies of the variance of, say, x_1 given Q . Now, what is the variance of x_1 given Q ?

Well, x_1 is just a Bernoulli random variable. But the difference is that for x , we don't know what the bias or what the Q is. Because it's some random bias Q .

But just like we said earlier in part A, when we talked about the expectation of x_1 given Q , this is actually just Q times 1 minus Q . Because if you knew what the bias were, it would be p times 1 minus p . So the bias times 1 minus the bias.

But you don't know what it is. But if you did, it would just be q . So what we do is we just plug in Q , and you get Q times 1 minus 2.

All right, and now this is expectation of n . I can pull out the n . So it's n times the expectation of Q minus Q^2 , which is just n times expectation Q , we can use linearity of expectations again, expectation of Q is μ .

And the expectation of Q^2 is, well, we can do that on the side. Expectation of Q^2 is the variance of Q plus expectation of Q quantity squared. So that's just σ^2 plus μ^2 . And so this is just going to be then minus σ^2 minus μ^2 .

All right, so that's the first term. Now let's do the second term. The variance the conditional expectation of x given Q . And again, what we can do is we can write x as the sum of all these x_i 's.

And now we can apply linearity of expectations. So we would get n times one of these expectations. And remember, we said earlier the expectation of x_1 given Q is just Q . So it's the variance of n times Q .

And remember now, n is just-- it's not random. It's just some number. So when you pull it out of a variance, you square it. So this is n squared times the variance of Q .

And the variance of Q we're given is σ^2 . So this is n squared times σ^2 . So the final answer is just a combination of these two terms. This one and this one.

So let's write it out. The variance of x , then, is equal to-- we can combine terms a little bit. So the first one, let's take the mus and we'll put them together. So it's $n\mu - \mu^2$.

And then we have n squared times σ^2 from this term and minus n times σ^2 from this term. So it would be n squared minus n times σ^2 , or n times n minus 1 times σ^2 . So that is the final answer that we get for the variance of x .

And now, let's try doing it another way. So that's one way of doing it. That's using the law of total expectations and conditioning on Q . Another way of finding the variance of x is to use the formula involving covariances, right? And we can use that because x is actually a sum of multiple random variables x_1 through x_n .

And the formula for this is, you have n variance terms plus all these other ones. Where i is not equal to j , you have the covariance terms. And really, it's just, you can think of it as a double sum of all pairs of x_i and x_j where if i and j happen just to be the same, that it simplifies to be just the variance. Now, so we pulled these n terms out because they are different than these because they have a different value.

And now fortunately, we've already calculated what these values are in part B. So we can just plug them them. All the variances are the same. And there's n of them, so we get n times the variance of each one. The variance of each one we calculated already was $\mu - \mu^2$.

And then, we have all the terms were i is not equal to j . Well, there are actually n squared minus n of them. So because you can take any one of the n 's to be the first to be i , any one of the n to be j . So that gives you n squared pairs.

But then you have to subtract out all the ones where i and j are the same. And there are n of them. So that leaves you with n squared minus n of these pairs where i is not equal to j .

And the coherence for this case where i is not equal to j , we also calculated in part B. That's just σ^2 . All right, and now if we compare these two, we'll see that they are proportionally

exactly the same. So we've use two different methods to calculate the variance, one using this summation and one using the law of total variance.

So what do we learn from this problem? Well, we saw that first of all, in order to find some expectations, it's very useful to use law of iterated expectations. But the trick is to figure out what you should condition on. And that's kind of an art that you learn through more practice.

But one good rule of thumb is, when you have kind of a hierarchy or layers of randomness where one layer of randomness depends on the randomness of the layer above-- so in this case, whether or not you get heads or tails depends on, that's random, but that depends on the randomness on the level above, which was the random bias of the coin itself. So the rule of thumb is, when you want to calculate the expectations for the layer where you're talking about heads or tails, it's useful to condition on the layer above where that is, in this case, the random bias. Because once you condition on the layer above, that makes the next level much simpler. Because you kind of assume that you know what all the previous levels of randomness are, and that helps you calculate what the expectation for this current level. And the rest of the problem was just kind of going through exercises of actually applying the--

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013 Transcript – Tutorial:A Random Number of Coin Flips

Hey, everyone. Welcome back. Today, we're going to do another fun problem that has to do with a random number of coin flips. So the experiment we're going to run is as follows. We're given a fair six-sided die, and we roll it.

And then we take a fair coin, and we flip it the number of times indicated by the die. That is to say, if I roll a four on my die, then I flip the coin four times. And then we're interested in some statistics regarding the number of heads that show up in our sequence. In particular, we want to compute the expectation and the variance of the number of heads that we see.

So the first step of this problem is to translate the English to the math. So we have to define some notation. I went ahead and did that for us. I defined n to be the outcome of the die role. Now, since we flip the coin the number of times shown by the die roll, n is equivalently the number of flips that we perform. And n , of course, is a random variable, and I've written its PMF up here.

So P_n of n is just a discrete uniform random variable between 1 and 6, because we're told that the die has six sides and that it's fair. Now, I also defined h to be the number of heads that we see. So that's the quantity of interest. And it turns out that Bernoulli random variables will be very helpful to us in this problem.

So I defined $x_{\text{sub } i}$ as 1 if the i th flip is heads, and 0 otherwise. And what we're going to do now is, we're going to use these $x_{\text{sub } i}$'s to come up with an expression for h . So if you want to count the number of heads, one possible thing you could do is start with 0 and then look at the first coin flip. If it's heads, you add 1 to 0, which I'm going to call your running sum.

If the first flip is tails, you add 0. And similarly, after that, after every trial, if you see heads, you add 1 to your running sum. If you see a tails, you add 0. And in that way, we can precisely compute h . So the mathematical statement of what I just said is that h is equal to x_1 plus x_2 plus x_3 , all the way through $x_{\text{sub } n}$.

So now, we are interested in computing e of h , the expectation of h . So your knee jerk reaction might be to say, oh, well, by linearity of expectation, we know that this is an expectation of x_1 , et cetera through the expectation of x_n . But in this case, you would actually be wrong. Don't do that.

And the reason that this is not going to work for us is because we're dealing with a random number of random variables. So each x_i is a random variable. And we have capital n of them. But capital n is a random variable. It denotes the outcome of our die roll.

So we actually cannot just take the sum of these expectations. Instead, we're going to have to condition on n and use iterated expectation. So this is the mathematical statement of what I just said. And the reason why this works is because conditioning on n will take us to the case that we

already know how to deal with, where we have a known number of random variables. And of course, iterated expectations holds, as you saw in lecture.

I will briefly mention here that the formula we're going to derive is derived in the book. And it was probably derived in lecture. So if you want, you can just go to that formula immediately. But I think the derivation of the formula that we need is quick and is helpful. So I'm going to go through it quickly.

Let's do it over here. Plugging in our running sum for h , we get this expression-- x_1 plus x_2 et cetera plus x_n , conditioned on n . And this, of course, is n times the expectation of $x_{\text{sub } i}$. So again, I'm going through this quickly, because it's in the book. But this step holds, because each of these x_i 's have the same statistics.

They're all Bernoulli with parameter of $1/2$, because our coin is fair. And so I used $x_{\text{sub } i}$ to say it doesn't really matter which integer you pick for i , because the expectation of x_i is the same for all i . So this now, the expectation of $x_{\text{sub } i}$, this is just a number, it's just some constant, so you can pull it out of the expectation. So you get the expectation of $x_{\text{sub } i}$ times the expectation of n .

So I gave away the answer to this a second ago. But $x_{\text{sub } i}$ is just a Bernoulli random variable with parameter of success of $1/2$. And we know already that the expectation of such a random variable is just p , or $1/2$. So this is $1/2$ times expectation of n . And now n we know is a discrete uniform random variable.

And there's a formula that I'm going to use, which hopefully some of you may remember. If you have a discrete uniform random variable that takes on values between a and b -- let's use w -- if you call this random variable w , then we have that the variance of w is equal to b minus a times b minus a plus 2 divided by 12. So that's the variance.

We don't actually need the variance, but we will need this later. And the expectation of w -- actually, let's just do it up here right ahead for this problem. Because we have a discrete uniform random variable, the expectation is just the middle. So you agree hopefully that the middle is right at 3.5, which is also $7/2$. So this is times $7/2$, which is equal to $7/4$.

So we are done with part of part a. I'm going to write this answer over here, so I can erase. And we're going to do something very similar to compute the variance. To compute the variance, we are going to also condition on n . So we get rid of this source of randomness. And then we're going to use law of total variance, which you've also seen in lecture. And again, the formula for this variance is derived in the book.

So I'm going to go through it quickly. But make sure you understand this derivation, because it exercises a lot of stuff we taught you. So this, just using law of total variance, is the variance of expectation of h given n , plus the expectation of the variance of h given n . And now, plugging in this running sum for h , you get this. It's a mouthful to write.

Bear with me. x_1 through x_n given n -- so I didn't do anything fancy. I just plugged this into here. So this term is similar to what we saw in a previous problem. By linearity of expectation and due

to the fact that all of the x_i 's are distributed in the same way, they have the same expectation, this becomes n times the expectation of x_i . And let's do this term over here.

This term-- well, conditioned on n , this n is known. So we essentially have a finite known sum of independent random variables. We know that the variance of a sum of independent random variables is the sum of the variances.

So this is the variance of x_1 plus the variance of x_2 et cetera, plus the variance of x_n . And furthermore, again, because all of these x_i 's have the same distribution, the variance is the same. So we can actually write this as n times the variance of x_i , where x_i just corresponds to one of the trials. It doesn't matter which one, because they all have the same variance and expectation.

So now, we're almost home free. This is just some scalar. So we can take it out of the variance, but we have to square it. So this becomes expectation of x_i^2 times the variance of n . And then this variance is also just a scalar, so we can take it outside.

So then we get variance of x_i times expectation of n . Now, we know that the expectation of x_i is just the probability of success, which is $1/2$. So we have $1/2$ squared, or $1/4$, times the variance of n . So that's where this formula comes in handy.

b is equal to 6, a is equal to 1. So we get that the variance of n is equal to 5 times-- and then 5 plus 2 is 7-- divided by 12. So this is just a formula from the book that you guys hopefully remember. So we get $35/12$. And then the variance of x_i , we know the variance of a Bernoulli random variable is just p times $1 - p$.

So in our case, that's $1/2$ times $1/2$, which is $1/4$. So we get $1/4$. And then the expectation of n , we remember from our previous computation, is just $7/2$. So I will let you guys do this arithmetic on your own time. But the answer comes out to be $77/48$.

So I will go ahead and put our answer over here-- $77/48$ -- so that I can erase. So I want you guys to start thinking about part b while I erase. Essentially, you do the same experiment that we did in part a, except now we use two dice instead of one. So in part b, just to repeat, you now have two dice.

You roll them. You look at the outcome. If you have an outcome of four on one die and six on another die, then you flip the coin 10 times. So it's the same exact experiment. We're interested in the number of heads we want the expectation and the variance. But this step is now a little bit different.

Again, let's approach this by defining some notation first. Now, I want to let n_1 be the outcome of the first die. And then you can let n_2 be the outcome of the second die. And we'll start with just that.

So one way you could approach this problem is say, OK, if n_1 is the outcome of my first die and n_2 is the outcome of my second die, then the number of coin flips that I'm going to make is n_1

plus n_2 . This is the total coin flips. So you could just repeat the same exact math that we did in part a, except everywhere that you see an n , you replace that n with n_1 plus n_2 .

So that will get you to your answer, but it will require slightly more work. We're going to think about this problem slightly differently. So the way we are thinking about it just now, we roll two dice at the same time. We add the results of the die rolls. And then we flip the coin that number of times.

But another way you can think about this is, you roll one die, and then you flip the coin the number of times shown by that die and count the number of heads. And then you take the second die and you roll it. And then you flip the coin that many more times and count the number of heads after that. So you could define h_1 to be number of heads in the first n_1 coin flips.

And you could just let h_2 be the number of heads in the last n_2 coin flips. So hopefully that terminology is not confusing you. Essentially, what I'm saying is, n_1 plus n_2 means you'll have n_1 flips, followed by n_2 flips, for a total of n_1 plus n_2 flips. And then within the first n_1 flips, you can get some number of heads, which we're calling h_1 .

And in the last n_2 flips, you can get some number of heads, which is h_2 . So the total number of heads that we get at the end-- I'm going to call it h^* -- is equal to h_1 plus h_2 . And what part b is really asking us for is the expectation of h^* and the variance of h^* . But here's where something really beautiful happens.

h_1 and h_2 are independent, and they are statistically the same. So the reason why they're independent is because-- well, first of all, all of our coin flips are independent. And they're statistically the same, because the experiment is exactly the same. And everything's independent.

So instead of imagining one person rolling two die and then summing the outcomes and flipping a coin that many times and counting heads, you can imagine one person takes one die and goes into one room. A second person takes a second die and goes into another room. They run their experiments.

Then they report back to a third person the number of heads. And that person adds them together to get h^* . And in that scenario, everything is very clearly independent. So the expectation of h^* -- you actually don't need independence for this part, because linearity of expectation always holds. But you get the expectation of h_1 plus the expectation of h_2 .

And because these guys are statistically equivalent, this is just two times the expectation of h . And the expectation of h we calculated in part a. So this is 2 times 7 over 4. Now, for the variance, here's where the independence comes in. I'm actually going to write this somewhere where I don't have to bend over.

So the variance of h^* is equal to the variance of h_1 plus the variance of h_2 by independence. And that's equal to 2 times the variance of h , because they are statistically the same. And the variance of h we computed already. So this is just 2 times 77 over 48.

So the sufficient answer to part b is that both the mean and the variance double from part A. So hopefully you guys enjoyed this problem. We covered a bunch of things. So we saw how to deal with having a random number of random variables.

Usually we have a fixed number of random variables. In this problem, the number of random variables we were adding together was itself random. So to handle that, we conditioned on n . And to compute expectation, we use iterated expectation.

To compute variance, we used law of total variance. And then in part b, we were just a little bit clever. We thought about how can we reinterpret this experiment to reduce computation. And we realized that part b is essentially two independent trials of part a. So both the mean and the variance should double.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013 Transcript – Lecture 12

The following content is provided under a Creative Commons license. Your support will help MIT OpenCourseWare continue to offer high quality educational resources for free. To make a donation or view additional materials from hundreds of MIT courses, visit MIT OpenCourseWare at ocw.mit.edu.

JOHN TSITSIKLIS: So today we're going to finish with the core material of this class. That is the material that has to do with probability theory in general. And then for the rest of the semester we're going to look at some special types of models, talk about inference. Well, there's also going to be a small module of core material coming later.

But today we're basically finishing chapter four. And what we're going to do is we're going to look at a somewhat familiar concept, the concept of the conditional expectation. But we're going to look at it from a slightly different angle, from a slightly more sophisticated angle. And together with the conditional expectation we will also talk about conditional variances. It's something that we're going to denote this way.

And we're going to see what they are, and there are some subtle concepts that are involved here. And we're going to apply some of the tools we're going to develop to deal with a special type of situation in which we're adding random variables. But we're adding a random number of random variables.

OK, so let's start talking about conditional expectations. I guess you know what they are. Suppose we are in the discrete world. xy , or discrete random variables. We defined the conditional expectation of x given that I told you the value of the random variable y . And the way we define it is the same way as an ordinary expectation, except that we're using the conditional PMF.

So we're using the probabilities that apply to the new universe where we are told the value of the random variable y . So this is still a familiar concept so far. If we're dealing with the continuous random variable x the formula is the same, except that here we have an integral, and we have to use the conditional density function of x .

Now what I'm going to do, I want to introduce it gently through the example that we talked about last time. So last time we talked about having a stick that has a certain length. And we take that stick, and we break it at some point that we choose uniformly at random. And let's denote why the place where we chose to break it.

Having chosen y , then we're left with a piece of the stick. And I'm going to choose a place to break it once more uniformly at random between 0 and y . So this is the second place at which we are going to break it, and we call that place x .

OK, so what's the conditional expectation of x if I tell you the value of y ? I tell you that capital Y happens to take a specific numerical value. So this capital Y is now a specific numerical value, x is chosen uniformly over this range. So the expected value of x is going to be half of this range between 0 and y . So the conditional expectation is little y over 2.

The important thing to realize here is that this quantity is a number. I told you that the random variable took a certain numerical value, let's say 3.5. And then you tell me given that the random variable took the numerical value 3.5 the expected value of x is 1.75. So this is an equality between numbers.

On the other hand, before you do the experiment you don't know what y is going to turn out to be. So this little y is the numerical value that has been observed when you start doing the experiments and you observe the value of capital Y . So in some sense this quantity is not known ahead of time, it is random itself. So maybe we can start thinking of it as a random variable.

So to put it differently, before we do the experiment I ask you what's the expected value of x given y ? You're going to answer me well I don't know, it depends on what y is going to turn out to be. So the expected value of x given y itself can be viewed as a random variable, because it depends on the random variable capital Y .

So hidden here there's some kind of statement about random variables instead of numbers. And that statement about random variables, we write it this way. By thinking of the expected value, the conditional expectation, as a random variable instead of a number. It's a random variable when we do not specify a specific number, but we think of it as an abstract object.

The expected value of x given the random variable y is the random variable y over 2 no matter what capital Y turns out to be. So we turn and take a statement that deals with equality of two numbers, and we make it a statement that's an equality between two random variables. OK so this is clearly a random variable because capital Y is random. What exactly is this object? I didn't yet define it for you formally. So let's now give the formal definition of this object that's going to be denoted this way.

The conditional expectation of x given the random variable y is a random variable. Which random variable is it? It's the random variable that takes this specific numerical value whenever capital Y happens to take the specific numerical value little y . In particular, this is a random variable, which is a function of the random variable capital Y . In this instance, it's given by a simple formula in terms of capital Y . In other situations it might be a more complicated formula.

So again, to summarize, it's a random. The conditional expectation can be thought of as a random variable instead of something that's just a number. So in any specific context when you're given the value of capital Y the conditional expectation becomes a number. This is the realized value of this random variable. But before the experiment starts, before you know what capital Y is going to be, all that you can say is that the conditional expectation is going to be 1/2 of whatever capital Y turns out to be.

This is a pretty subtle concept, it's an abstraction, but it's a useful abstraction. And we're going to see today how to use it. All right, I have made the point that the conditional expectation, the random variable that takes these numerical values is a random variable. If it is a random variable this means that it has an expectation of its own. So let's start thinking what the expectation of the conditional expectation is going to turn out to be.

OK, so the conditional expectation is a random variable, and in general it's some function of the random variable y that we are observing. In terms of numerical values if capital Y happens to take a specific numerical value then the conditional expectation also takes a specific numerical value, and we use the same function to evaluate it. The difference here is that this is an equality of random variables, this is an equality between numbers.

Now if we want to calculate the expected value of the conditional expectation we're basically talking about the expected value of a function of a random variable. And we know how to calculate expected values of a function. If we are in the discrete case, for example, this would be a sum over all y 's of the function who's expected value we're taking times the probability that y takes on a specific numerical value.

OK, but let's remember what g is. So g is the numerical value of the conditional expectation of x with y . And now when you see this expression you recognize it. This is the expression that we get in the total expectation theorem. Did I miss something? Yes, in the total expectation theorem to find the expected value of x , we divide the world into different scenarios depending on what y happens. We calculate the expectation in each one of the possible worlds, and we take the weighted average.

So this is a formula that you have seen before, and you recognize that this is the expected value of x . So this is a longer, more detailed derivation of what I had written up here, but the important thing to keep in mind is the moral of the story, the punchline. The expected value of the conditional expectation is the expectation itself.

So this is just our total expectation theorem, but written in more abstract notation. And it comes handy to have this more abstract notation, as as we're going to see in a while. OK, we can apply this to our stick example. If we want to find the expected value of x how much of the stick is left at the end?

We can calculate it using this law of iterated expectations. It's the expected value of the conditional expectation. We know that the conditional expectation is y over 2. So expected value of y is 1 over 2, because y is uniform so we get 1 over 4. So this gives us the same answer that we derived last time in a rather long way.

All right, now that we have mastered conditional expectations, let's raise the bar a little more and talk about conditional variances. So the conditional expectation is the mean value, or the expected value, in a conditional universe where you're told the value of y . In that same conditional universe you can talk about the conditional distribution of x , which has a mean-- the conditional expectation-- but the conditional distribution of x also has a variance.

So we can talk about the variance of x in that conditional universe. The conditional variance as a number is the natural thing. It's the variance of x , except that all the calculations are done in the conditional universe. In the conditional universe the expected value of x is the conditional expectation. This is the distance from the mean in the conditional universe squared. And we take the average value of the squared distance, but calculate it again using the probabilities that apply in the conditional universe.

This is an equality between numbers. I tell you the value of y , once you know that value for y you can go ahead and plot the conditional distribution of x . And for that conditional distribution you can calculate the number which is the variance of x in that conditional universe. So now let's repeat the mental gymnastics from the previous slide, and abstract things, and define a random variable-- the conditional variance. And it's going to be a random variable because we leave the numerical value of capital Y unspecified.

So ahead of time we don't know what capital Y is going to be, and because of that we don't know ahead of time what the conditional variance is going to be. So before the experiment starts if I ask you what's the conditional variance of x ? You're going to tell me well I don't know, It depends on what y is going to turn out to be. It's going to be something that depends on y . So it's a random variable, which is a function of y .

So more precisely, the conditional variance when written in this notation just with capital letters, is a random variable. It's a random variable whose value is completely determined once you learned the value of capital Y . And it takes a specific numerical value. If capital Y happens to get a realization that's a specific number, then the variance also becomes a specific number. And it's just a conditional variance of y over x in that universe.

All right, OK, so let's continue what we did in the previous slide. We had the law of iterated expectations. That told us that expected value of a conditional expectation is the unconditional expectation. Is there a similar rule that might apply in this context?

So you might guess that the variance of x could be found by taking the expected value of the conditional variance. It turns out that this is not true. There is a formula for the variance in terms of conditional quantities. But the formula is a little more complicated. If involves two terms instead of one. So we're going to go quickly through the derivation of this formula. And then, through examples we'll try to get some interpretation of what the different terms here correspond to.

All right, so let's try to prove this formula. And the proof is sort of a useful exercise to make sure you understand all the symbols that are involved in here. So the proof is not difficult, it's 4 and 1/2 lines of algebra, of just writing down formulas. But the challenge is to make sure that at each point you understand what each one of the objects is.

So we go into formula for the variance affects. We know in general that the variance of x has this nice expression that we often use to calculate it. The expected value of the squared of the random variable minus the mean squared. This formula, for the variances, of course it should apply to

conditional universes. I mean it's a general formula about variances. If we put ourselves in a conditional universe where the random variable y is given to us the same math should work.

So we should have a similar formula for the conditional variances. It's just the same formula, but applied to the conditional universe. The variance of x in the conditional universe is the expected value of x squared-- in the conditional universe-- minus the mean of x -- in the conditional universe-- squared. So this formula looks fine.

Now let's take expected values of both sides. Remember the conditional variance is a random variable, because its value depends on whatever realization we get for capital Y . So we can take expectations here. We get the expected value of the variance.

Then we have the expected value of a conditional expectation. Here we use the fact that we discussed before. The expected value of a conditional expectation is the same as the unconditional expectation. So this term becomes this. And finally, here we just have some weird looking random variable, and we take the expected value of it.

All right, now we need to do something about this term. Let's use the same rule up here to write down this variance. So variance of an expectation, that's kind of strange, but you remember that the conditional expectation is random, because y is random. So this thing is a random variable, so this thing has a variance.

What is the variance of this thing? It's the expected value of the thing squared minus the square of the expected value of the thing. Now what's the expected value of that thing? By the law of iterated expectations, once more, the expected value of this thing is the unconditional expectation. And that's why here I put the unconditional expectation.

So I'm using again this general rule about how to calculate variances, and I'm applying it to calculate the variance of the conditional expectation. And now you notice that if you add these two expressions c and d we get this plus that, which is this. It's equal to-- these two terms cancel, we're left with this minus that, which is the variance of x . And that's the end of the proof.

This one of those proofs that do not convey any intuition. This, as I said, it's a useful proof to go through just to make sure you understand the symbols. It starts to get pretty confusing, and a little bit on the abstract side. So it's good to understand what's going on.

Now there is intuition behind this formula, some of which is better left for later in the class when we talk about inference. The idea is that the conditional expectation you can interpret it as an estimate of the random variable that you are trying to-- an estimate of x based on measurements of y , you can think of these variances as having something to do with an estimation error. And once you start thinking in those terms an interpretation will come about. But again as I said this is better left for when we start talking about inference.

Nevertheless, we're going to get some intuition about all these formulas by considering a baby example where we're going to apply the law of iterated expectations, and the law of total variance. So the baby example is that we do this beautiful experiment of giving a quiz to a class

consisting of many sections. And we're interested in two random variables. So we have a number of students, and they're all allocated to sections. The experiment is that I pick a student at random, and I look at two random variables.

One is the quiz score of the randomly selected student, and the other random variable is the section number of the student that I have selected. We're given some statistics about the two sections. Section one has 10 students, section two has 20 students. The quiz average in section one was 90. Quiz average in section two was 60. What's the expected value of x ? What's the expected quiz score if I pick a student at random?

Well, each student has the same probability of being selected. I'm making that assumption out of the 30 students. I need to add the quiz scores of all of the students. So I need to add the quiz scores in section one, which is 90 times 10. I need to add the quiz scores in that section, which is 60 times 20. And we find that the overall average was 70. So this is the usual unconditional expectation.

Let's look at the conditional expectation, and let's look at the elementary version where we're talking about numerical values. If I tell you that the randomly selected student was in section one what's the expected value of the quiz score of that student? Well, given this information, we're picking a random student uniformly from that section in which the average was 90. The expected value of the score of that student is going to be 90.

So given the specific value of y , the specific section, the conditional expectation or the expected value of the quiz score is a specific number, the number 90. Similarly for the second section the expected value is 60, that's the average score in the second section. This is the elementary version. What about the abstract version? In the abstract version the conditional expectation is a random variable because it depends. In which section is the student that I picked?

And with probability 1/3, I'm going to pick a student in the first section, in which case the conditional expectation will be 90, and with probability 2/3 I'm going to pick a student in the second section. And in that case the conditional expectation will take the value of 60. So this illustrates the idea that the conditional expectation is a random variable. Depending on what y is going to be, the conditional expectation is going to be one or the other value with certain probabilities.

Now that we have the distribution of the conditional expectation we can calculate the expected value of it. And the expected value of such a random variable is 1/3 times 90, plus 2/3 times 60, and it comes out to equal 70. Which miraculously is the same number that we got up there.

So this tells you that you can calculate the overall average in a large class by taking the averages in each one of the sections and weighing each one of the sections according to the number of students that it has. So this section had 90 students but only 1/3 of the students, so it gets a weight of 1/3.

So the law of iterated expectations, once more, is nothing too complicated. It's just that you can calculate overall class average by looking at the section averages and combine them. Now since

the conditional expectation is a random variable, of course it has a variance of its own. So let's calculate the variance. How do we calculate variances? We look at all the possible numerical values of this random variable, which are 90 and 60. We look at the difference of those possible numerical values from the mean of this random variable, and the mean of that random variable, we found that's it's 70. And then we weight the different possible numerical values according to their probabilities.

So with probability 1/3 the conditional expectation is 90, which is 20 away from the mean. And we get this squared distance. With probability 2/3 the conditional expectation is 60, which is 10 away from the mean, has this squared distance and gets weighed by 2/3, which is the probability of 60. So you do the numbers, and you get the value for the variance equal to 200.

All right, so now we want to move towards using that more complicated formula involving the conditional variances. OK, suppose someone goes and calculates the variance of the quiz scores inside each one of the sections. So someone gives us these two pieces of information. In section one we take the differences from the mean in that section, and let's say that the various turns out to be a number equal to 10 similarly in the second section. So these are the variances of the quiz scores inside individual sections.

The variance in one conditional universe, the variance in the other conditional universe. So if I pick a student in section one and I don't tell you anything more about the student, what's the variance of the random score of that student? The variance is 10.

I know why, but I don't know the student. So the score is still a random variable in that universe. It has a variance, and that's the variance. Similarly, in the other universe, the variance of the quiz scores is this number, 20.

Once more, this is an equality between numbers. I have fixed the specific value of y . So I put myself in a specific universe, I can calculate the variance in that specific universe.

If I don't specify a numerical value for capital Y , and say I don't know what Y is going to be, it's going to be random. Then what kind of section variance I'm going to get itself will be random. With probability 1/3, I pick a student in the first section in which case the conditional variance given what I have picked is going to be 10. Or with probability 2/3 I pick y equal to 2, and I place myself in that universe. And in that universe the conditional variance is 20.

So you see again from here that the conditional variance is a random variable that takes different values with certain probabilities. And which value it takes depends on the realization of the random variable capital Y . So this happens if capital Y is one, this happens if capital Y is equal to 2.

Once you have something of this form-- a random variable that takes values with certain probabilities-- then you can certainly calculate the expected value of that random variable. Don't get intimidated by the fact that this random variable, it's something that's described by a string of eight symbols, or seven, instead of just a single letter. Think of this whole string of symbols there as just being a random variable. You could call it z for example, use one letter.

So Z is a random variable that takes these two values with these corresponding probabilities. So we can talk about the expected value of Z , which is going to be $1/3$ times 10 , $2/3$ times 20 , and we get a certain number from here.

And now we have all the pieces to calculate the overall variance of x . The formula from the previous slide tells us this. Do we have all the pieces? The expected value of the variance, we just calculated it. The variance of the expected value, this was the last calculation in the previous slide. We did get a number for it, it was 200 . You add the two, you find the total variance.

Now the useful piece of this exercise is to try to interpret these two numbers, and see what they mean. The variance of x given y for a specific y is the variance inside section one. This is the variance inside section two. The expected value is some kind of average of the variances inside individual sections.

So this term tells us something about the variability of this course, how widely spread they are within individual sections. So we have three sections, and this course happens to be-- OK, let's say the sections are really different. So here you have undergraduates and here you have post-doctoral students. And these are the quiz scores, that's section one, section two, section three. Here's the mean of the first section. And the variance has something to do with the spread. The variance in the second section has something to do with the spread, similarly with the third spread.

And the expected value of the conditional variances is some weighted average of the three variances that we get from individual sections. So variability within sections definitely contributes something to the overall variability of this course. But if you ask me about the variability over the entire class there's a second effect. That has to do with the fact that different sections are very different from each other. That these courses here are far away from those scores.

And this term is the one that does the job. This one looks at the expected values inside each section, and these expected values are this, this, and that. And asks a question how widely spread are they? It asks how different from each other are the means inside individual sections? And in this picture it would be a large number because the difference section means are quite different.

So the story that this formula is telling us is that the overall variability of the quiz scores consists of two factors that can be quantified and added. One factor is how much variability is there inside individual sections? And the other factor is how different are the sections from each other? Both effects contribute to the overall variability of this course.

Let's continue with just one more numerical example. Just to get the hang of doing these kinds of calculations, and apply this formula to do a divide and conquer calculation of the variance of a random variable. Just for variety now we're going to take a continuous random variable.

Somebody gives you a PDF if this form, and they ask you for the variance. And you say oh that's too complicated, I don't want to do integrals. Can I divide and conquer?

And you say OK, let me do the following trick. Let me define a random variable, y . Which takes the value 1 if x falls in here, and takes the value 2 if x falls in the second interval. And let me try to work in the conditional world where things might be easier, and then add things up to get the overall variance.

So I have defined y this particular way. In this example y becomes a function of x . y is completely determined by x . And I'm going to calculate the overall variance by trying to calculate all of the terms that are involved here. So let's start calculating.

First observation is that this event has probability $1/3$, and this event has probability $2/3$. The expected value of x given that we are in this universe is $1/2$, because we have a uniform distribution from 0 to 1. Here we have a uniform distribution from 1 to 2, so the conditional expectation of x in that universe is $3/2$.

How about conditional variances? In the world who are y is equal to 1 x has a uniform distribution on a unit interval. What's the variance of x ? By now you've probably seen that formula, it's $1/12$. $1/12$ is the variance of a uniform distribution over a unit interval.

When y is equal to 2 the variance is again $1/12$. Because in this instance again x has a uniform distribution over an interval of unit length. What's the overall expected value of x ? The way you find the overall expected value is to consider the different numerical values of the conditional expectation. And weigh them according to their probabilities.

So with probability $1/3$ the conditional expectation is $1/2$. And with probability $2/3$ the conditional expectation is $3/2$. And this turns out to be $7/6$.

So this is the advance work we need to do, now let's calculate a few things here. What's the variance of the expected value of x given y ? Expected value of x given y is a random variable that takes these two values with these probabilities.

So to find the variance we consider the probability that the expected value takes the numerical value of $1/2$ minus the mean of the conditional expectation. What's the mean of the conditional expectation? It's the unconditional expectation. So it's $7/6$. We just did that calculation. So I'm putting here that number, $7/6$ squared. And then there's a second term with probability $2/3$, the conditional expectation takes this value of $3/2$, which is so much away from the mean, and we get this contribution.

So this way we have calculated the variance of the conditional expectation, this is this term. What is this? Any guesses what this number is? It's $1/12$, why? The conditional variance just happened in this example to be $1/12$ no matter what. So the conditional variance is a deterministic random variable that takes a constant value. So the expected value of this random variable is just $1/12$. So we got the two pieces that we need, and so we do have the overall variance of the random variable x .

So this was just an academic example in order to get the hang of how to manipulate various quantities. Now let's use what we have learned and the tools that we have to do something a little

more interesting. OK, so by now you're all in love with probabilities. So over the weekend you're going to bookstores to buy probability books.

So you're going to visit a random number bookstores, and at each one of the bookstores you're going to spend a random amount of money. So let n be the number of stores that you are visiting. So n is an integer-- non-negative random variable-- and perhaps you know the distribution of that random variable.

Each time that you walk into a store your mind is clear from whatever you did before, and you just buy a random number of books that has nothing to do with how many books you bought earlier on the day. It has nothing to do with how many stores you are visiting, and so on. So each time you enter as a brand new person, and buy a random number of books, and spend a random amount of money.

So what I'm saying, more precisely, is that I'm making the following assumptions. That for each store i -- if you end up visiting the i -th store-- the amount of money that you spend is a random variable that has a certain distribution. That distribution is the same for each store, and the x_i 's from store to store are independent from each other. And furthermore, the x_i 's are all independent of n . So how much I'm spending at the store-- once I get in-- has nothing to do with how many stores I'm visiting.

So this is the setting that we're going to look at. y is the total amount of money that you did spend. It's the sum of how much you spent in the stores, but the index goes up to capital N . And what's the twist here? It's that we're dealing with the sum of independent random variables except that how many random variables we have is not given to us ahead of time, but it is chosen at random.

So it's a sum of a random number of random variables. We would like to calculate some quantities that have to do with y , in particular the expected value of y , or the variance of y . How do we go about it?

OK, we know something about the linearity of expectations. That expectation of a sum is the sum of the expectations. But we have used that rule only in the case where it's the sum of a fixed number of random variables. So expected value of x plus y plus z is expectation of x , plus expectation of y , plus expectation of z . We know this for a fixed number of random variables. We don't know it, or how it would work for the case of a random number.

Well, if we know something about the case for fixed random variables let's transport ourselves to a conditional universe where the number of random variables we're summing is fixed. So let's try to break the problem divide and conquer by conditioning on the different possible values of the number of bookstores that we're visiting. So let's work in the conditional universe, find the conditional expectation in this universe, and then use our law of iterated expectations to see what happens more generally.

If I told you that I visited exactly little n stores, where little n now is a number, let's say 10. Then the amount of money you're spending is x_1 plus x_2 all the way up to x_{10} given that we visited 10

stores. So what I have done here is that I've replaced the capital N with little n, and I can do this because I'm now in the conditional universe where I know that capital N is little n. Now little n is fixed.

We have assumed that n is independent from the x_i 's. So in this universe of a fixed n this information here doesn't tell me anything new about the values of the x 's. If you're conditioning random variables that are independent from the random variables you are interested in, the conditioning has no effect, and so it can be dropped.

So in this conditional universe where you visit exactly 10 stores the expected amount of money you're spending is the expectation of the amount of money spent in 10 stores, which is the sum of the expected amount of money in each store. Each one of these is the same number, because the random variables have identical distributions. So it's n times the expected value of money you spent in a typical store.

This is almost obvious without doing it formally. If I'm telling you that you're visiting 10 stores, what you expect to spend is 10 times the amount you expect to spend in each store individually. Now let's take this equality here and rewrite it in our abstract notation, in terms of random variables. This is an equality between numbers. Expected value of y given that you visit 10 stores is 10 times this particular number.

Let's translate it into random variables. In random variable notation, the expected value of money you're spending given the number of stores-- but without telling you a specific number-- is whatever that number of stores turns out to be times the expected value of x . So this is a random variable that takes this as a numerical value whenever capital N happens to be equal to little n. This is a random variable, which by definition takes this numerical value whenever capital N is equal to little n.

So no matter what capital N happens to be what specific value, little n, it takes this is equal to that. Therefore the value of this random variable is going to be equal to that random variable. So as random variables, these two random variables are equal to each other.

And now we use the law of iterated expectations. The law of iterated expectations tells us that the overall expected value of y is the expected value of the conditional expectation. We have a formula for the conditional expectation. It's n times expected value of x .

Now the expected value of x is a number. Expected value of something random times a number is expected value of the random variable times the number itself. We can take a number outside the expectation.

So expected value of x gets pulled out. And that's the conclusion, that overall the expected amount of money you're going to spend is equal to how many stores you expect to visit on the average, and how much money you expect to spend on each one on the average. You might have guessed that this is the answer. If you expect to visit 10 stores, and you expect to spend \$100 on each store, then yes, you expect to spend \$1,000 today. You're not going to impress your Harvard friends if you tell them that story.

It's one of the cases where reasoning, on the average, does give you the plausible answer. But you will be able to impress your Harvard friends if you tell them that I can actually calculate the variance of how much I can spend. And we're going to work by applying this formula that we have, and the difficulty is basically sorting out all those terms here, and what they mean.

So let's start with this term. So the expected value of y given that you're visiting n stores is n times the expected value of x . That's what we did in the previous slide. So this thing is a random variable, it has a variance. What is the variance? Is the variance of n times the expected value of x .

Remember expected value of x is a number. So we're dealing with the variance of n times a number. What happens when you multiply a random variable by a constant? The variance becomes the previous variance times the constant squared. So the variance of this is the variance of n times the square of that constant that we had here.

So this tells us the variance of the expected value of y given n . This is the part of the variability of how much money you're spending, which is attributed to the randomness, or the variability, in the number of stores that you are visiting. So the interpretation of the two terms is there's randomness in how much you're going to spend, and this is attributed to the randomness in the number of stores together with the randomness inside individual stores. Well, after I tell you how many stores you're visiting.

So now let's deal with this term-- the variance inside individual stores. Let's take it slow. If I tell you that you're visiting exactly little n stores, then y is how much money you spent in those little n stores. You're dealing with the sum of little n random variables. What is the variance of the sum of little n random variables? It's the sum of their variances.

So each store contributes a variance of x , and you're adding over little n stores. That's the variance of money spent if I tell you the number of stores. Now let's translate this into random variable notation. This is a random variable that takes this numerical value whenever capital N is equal to little n . This is a random variable that takes this numerical value whenever capital N is equal to little n . This is equal to that. Therefore, these two are always equal, no matter what happens to y .

So we have an equality here between random variables. Now we take expectations of both. Expected value of the variance is expected value of this. OK it may look confusing to think of the expected value of the variance here, but the variance of x is a number, not a random variable. You think of it as a constant. So its expected value of n times a constant gives us the expected value of n times the constant itself.

So now we got the second term as well, and now we put everything together, this plus that to get an expression for the overall variance of y . Which again, as I said before, the overall variability in y has to do with the variability of how much you spent inside the typical store. And the variability in the number of stores that you are visiting.

OK, so this is it for today. We'll change subjects quite radically from next time.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013 Transcript – Recitation: The Variance in the Stick Breaking Problem

Hi. In this problem, we'll get a chance to see the usefulness of conditioning in helping us to calculate quantities that would otherwise be difficult to calculate. Specifically, we'll be using the law of iterated expectations and the law of total variance.

Before we get started, let's just take a quick moment to interpret what these two laws are saying. Really, what it's saying is, in order to calculate the expectation or the variance of some random variable x , if that's difficult to do, we'll instead attack this problem in stages.

So the first stage is, we'll condition on some related random variable, y . And the hope is that by conditioning on this and reducing it to this conditional universe, the expectation of x will be easier to calculate.

Now, recall that this conditional expectation is really a random variable, which is a function of the random variable y . So what we've done is we first average out x given some y . What remains is some new random variable, which is a function of y . And now, what we have is randomness in y , which will then average out again to get the final expectation of x .

OK, so in this problem, we'll actually see an example of how this plays out. One more thing before we get started that's useful to recall is if y is a uniform random variable, distributed between a and b , then the variance of y is b minus a squared over 12, and the expectation of y is just a midpoint, a plus b over 2.

All right, so let's get started on the problem. So what we have is we have a stick of some fixed length, l , and what we do is we break it uniformly at random. So what we do is we choose a point uniformly at random along this stick. And we break it there, and then we keep the left portion of that stick.

So let's call the length of this left portion after the first break random variable y . So it's random because the point where we break it is random. And then what we do is we repeat this process.

We'll take this left side of the stick that's left. And we'll pick another point, uniformly at random, along this left remaining side. And we'll break it again, and keep the left side of that break. And we'll call that the length of the final remaining piece, x , which again is random.

The problem is really asking us to calculate the expectation of variance of x . So at first, it seems difficult to do, because the expectation and variance of x depends on where you break it the second time and also where you break it the first time. So let's see if conditioning can help us here.

So the first thing that we'll notice is that, if we just consider y , the length of the stick after the first break, it's actually pretty easy to calculate the expectation and variance of y . Because y ,

when you think about it, is actually just the simple uniform in a variable, uniformly distributed between 0 and 1, the length of the stick.

And this is because we're told that we choose the point of the break uniformly at random between 0 and 1. And so wherever we choose it, that's going to be the length of the left side of the stick. And so because of this, we know that the expectation of y is just $1/2$, and the variance of y is 1 squared over 12 .

But unfortunately, calculating the expectation variance of x is not quite as simple, because x isn't just uniformly distributed between 0 and some fixed number. Because it's actually uniformly distributed between 0 and y , wherever the first break was. But where the first break is is random too. And so we can't just say that x is a uniformly distributed random variable.

So what do we do instead? Well, we'll make the nice observation that let's pretend that we actually know what y is. If we knew what y was, then calculating the expectation of x would be simple, right? So if we were given that y is just some little y , then x would in fact just be uniformly distributed between 0 and little y .

And then if that's the case, then our calculation is simple, because the expectation of x would just be $y/2$, and the variance would just be y squared over 12 . All right, so let's make that a little bit more formal. What we're saying is that the expectation of x , If we knew what y was, would just be $y/2$. And the variance of x If we knew what y was would just be y squared over 12 .

All right, so notice what we've done. We've taken the second stage and we've said, let's pretend we know what happens in the first stage where we break it. And we know what y , the first break, was. Then the second stage becomes simple, because the average of x is just going to be the midpoint.

Now what we do to calculate the actual expectation of x , well, we'll invoke the law of iterated expectations. So expectation of x is expectation of the conditional expectation of x given 1 , which in this case is just expectation of $y/2$. And we know what the expectation of y is. It's $1/2$. And so this is just $1/4$. $1/4$.

All right, and so notice what we've done. We've taken this calculation and done it in stages. So we assume we know where the first break is. Given that, the average location of the second break becomes simple. It's just in the midpoint.

And then, we move up to the higher stage. And that now we average out over where the first break could have been. And that gives us our final answer. And notice that this actually makes sense, if we just think about it intuitively, because on average, the first break will be somewhere in the middle. And then that will leave us with half the stick left, and we break it again. On average, that will leave us with another half. So on average, you get a quarter of the original stick left, which makes sense.

All right, so that's the first part, where we use the law of iterated expectations. Now, let's go to part B, where we're actually asked to find the variance.

The variance is given by the law of total variance. So let's do it in stages. We'll first calculate the first term, the expectation of the conditional variance. Well, what is the expectation of the conditional variance? We've already calculated out what this conditional variance is. The conditional variance is y^2 over 12. So let's just plug that in. It's expectation of y^2 over 12.

All right, now this looks like it could be a little difficult to calculate. But let's just first pull out the $1/12$. And then remember, one way to calculate the expectation of the square of a random variable is to use the variance.

So recall that the variance of any random variable is just expectation of the square minus the square of the expectation. So if we want to calculate the expectation of the square, we can just take the variance and add the square of the expectation.

So this actually we can get pretty easily. It's actually just the variance of y plus the square of the expectation of y . And we know what these two terms are. The variance of y is 1^2 over 12. And the expectation of y is $1/2$. So when you square that, you get 1^2 over 4. So 1^2 over 12 plus 1^2 over 4 gives you 1^2 over 3. And you get that the first term is 1^2 over 36.

All right, now let's calculate the second term. Second term is the variance of the conditional expectation. So the variance of expectation of x given y . Well, what is the expectation of x given y ? We've already calculated that. That's $y/2$. So what we really want is the variance of $y/2$.

And remember, when you have a constant inside the variance, you pull it out but you square it. So what you get is $1/4$ the variance of y , which we know that the variance of y is 1^2 over 12. So we get that this is 1^2 over 48.

OK, so we've calculated both terms of this conditional variance. So all we need to do to find the final answer is just to add them. So it's 1^2 over 36 plus 1^2 over 48. And so, the final answer is 7^2 over 144.

OK, and so this is the first, the expectation of x , maybe you could have guessed intuitively. But the variance of x is not something that looks like something that you could have calculated off the top of your head. And so I guess the lesson from this example is that it is often very helpful if you condition on some things, because it allows you to calculate things in stages and build up from the bottom.

But it's important to note that the choice of what you condition on-- so the choice of y -- is actually very important, because you could choose lots of other y 's that wouldn't actually help you at all. And so how to actually choose this y is something that you can learn based on just having practiced with these kinds of problems.

So again, the overall lesson is, conditioning can often help when you calculate these problems. And so you should look to see if that could be a possible solution. So I hope that was helpful, and see you next time.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013 Transcript – Tutorial: Using the Conditional Expectation and Variance

Hey guys. Welcome back. Today we're going to do a fun problem that will test your knowledge of the law of total variance. And in the process, we'll also get more practice dealing with joint PDFs and computing conditional expectations and conditional variances.

So in this problem, we are given a joint PDF for x and y . So we're told that x and y can take on the following values in the shape of this parallelogram, which I've drawn. And moreover, that x and y are uniformly distributed. So the joint PDF is just flat over this parallelogram. And because the parallelogram has an area of 1, the height of the PDF must also be 1 so that the PDF integrates to 1. OK.

And then we are asked to compute the variance of x plus y . So you can think of x plus y as a new random variable whose variance we want to compute. And moreover, we're told we should compute this variance by using something called the law of total variance. So from lecture, you should remember or you should recall that the law of total variance can be written in these two ways.

And the reason why there's two different forms for this case is because the formula always has you conditioning on something. Here we condition on x , here we condition on y . And for this problem, the logical choice you have for what to condition on is x or y . So again, we have this option. And my claim is that we should condition on x . And the reason has to do with the geometry of this diagram.

So notice that if you freeze an x and then you sort of vary x , the width of this parallelogram stays constant. However, if you condition on y and look at the width this way, you see that the width of the slices you get by conditioning vary with y . So to make our lives easier, we're going to condition on x . And I'm going to erase this bottom one, because we're not using it.

So this really can seem quite intimidating, because we have nested variances and expectations going on, but we'll just take it slowly step by step. So first, I want to focus on this term-- the conditional expectation of x plus y conditioned on x . So coming back over to this picture, if you fix an arbitrary x in the interval, 0 to 1, we're restricting ourselves to this universe. So y can only vary between this point and this point.

Now, I've already written down here that the formula for this line is given by y is equal to x . And the formula for this line is given by y is equal to x plus 1. So in particular, when we condition on x , we know that y varies between x and x plus 1. But we actually know more than that. We know that in the unconditional universe, x and y were uniformly distributed. So it follows that in the conditional universe, y should also be uniformly distributed, because conditioning doesn't change the relative frequency of outcomes.

So that reasoning means that we can draw the conditional PDF of y conditioned on x as this. We said it varies between x and x plus 1. And we also said that it's uniform, which means that it must have a height of 1. So this is $p(y \text{ given } x)$.

Now, you might be concerned, because, well, we're trying to compute the expectation of x plus y and this is the conditional PDF of y , not of the random variable, x plus y . But I claim that we're OK, this is still useful, because if we're conditioning on x , this x just acts as a constant. It's not really going to change anything except shift the expectation of y by an amount of x . So what I'm saying in math terms is that this is actually just x plus the expectation of y given x .

And now our conditional PDF comes into play. Conditioned on x , this is the PDF of y . And because it's uniformly distributed and because expectation acts like center of mass, we know that the expectation should be the midpoint, right? And so to compute this point, we simply take the average of the endpoints, x plus 1 plus x over 2, which gives us $2x$ plus 1 over 2. So plugging this back up here, we get $2x/2$ plus $2x$ plus 1 over 2, which is $4x$ plus 1 over 2, or $2x$ plus $1/2$. OK.

So now I want to look at the next term, the next inner term, which is this guy. So this computation is going to be very similar in nature, actually. So we already discussed that the joint-- sorry, not the joint, the conditional PDF of y given x is this guy. So the variance of x plus y conditioned on x , we sort of have a similar phenomenon occurring. x now in this conditional world just acts like a constant that shifts the PDF but doesn't change the width of the distribution at all. So this is actually just equal to the variance of y given x , because constants don't affect the variance.

And now we can look at this conditional PDF to figure out what this is. So we're going to take a quick tangent over here, and I'm just going to remind you guys that we have a formula for computing the variance of a random variable when it's uniformly distributed between two endpoints. So say we have a random variable whose PDF looks something like this.

Let's call it, let's say, w . This is $p(w)$. We have a formula that says variance of w is equal to b minus a squared over 12. So we can apply that formula over here. b is x plus 1, a is x . So b minus a squared over 12 is just $1/12$. So we get $1/12$. So we're making good progress, because we have this inner quantity and this inner quantity.

So now all we need to do is take the outer variance and the outer expectation. So writing this all down, we get variance of x plus y is equal to variance of this guy, $2x$ plus $1/2$ plus the expectation of $1/12$. So this term is quite simple. We know that the expectation of a constant or of a scalar is simply that scalar. So this evaluates to $1/12$.

And this one is not bad either. So similar to our discussion up here, we know constants do not affect variance. You know they shift your distribution, they don't change the variance. So we can ignore the $1/2$. This scaling factor of 2, however, will change the variance. But we know how to handle this already from previous lectures. We know that you can just take out this scalar scaling factor as long as we square it. So this becomes 2^2 or 4 times the variance of x plus $1/12$.

And now to compute the variance of x , we're going to use that formula again, and we're going to use this picture. So here we have the joint PDF of x and y , but really we want now the PDF of x , so we can figure out what the variance is. So hopefully you remember a trick we taught you called marginalization. To get the PDF of x given a joint PDF, you simply marginalize over the values of y .

So if you freeze x is equal to 0, you get the probability density line over x by integrating over this interval, over y . So if you integrate over this strip, you get 1. If you move x over a little bit and you integrate over this strip, you get 1. This is the argument I was making earlier that the width of this interval stays the same, and hence, the variance stays the same. So based on that argument, which was slightly hand wavy, let's come over here and draw it.

We're claiming that the PDF of x , p_x of x , looks like this. It's just uniformly distributed between 0 and 1. And if you buy that, then we're done, we're home free, because we can apply this formula, b minus a squared over 12, gives us the variance. So b is 1, a is 0, which gives variance of x is equal to $1/12$. So coming back over here, we get 4 times $1/12$ plus $1/12$, which is $5/12$. And that is our answer.

So this problem was straightforward in the sense that our task was very clear. We had to compute this, and we had to do so by using the law of total variance. But we sort of reviewed a lot of concepts along the way. We saw how, given a joint PDF, you marginalize to get the PDF of x . We saw how constants don't change variance. We got a lot of practice finding conditional distributions and computing conditional expectations and variances.

And we also saw this trick. And it might seem like cheating to memorize formulas, but there's a few important ones you should know. And it will help you sort of become faster at doing computations. And that's important, especially if you guys take the exams. So that's it. See you next time.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013
Transcript – Recitation: Widgets and Crates

Hi. In this problem, we'll get more practice using conditioning to help us calculate expectations of variances. We'll see that in this problem, which deals with widgets and crates, it's actually similar in flavor to an earlier problem that we did, involving breaking a stick twice. And you'll see that in this problem, we'll again use the law of iterated expectations and the law of total variance to help us calculate expectations of variances.

And again, we'll be taking the approach of attacking the problem by splitting into the stages and building up from the bottom up. So in this problem, what we have is a crate, which contains some number of boxes. And we don't know how many boxes are. It's random. And it's given by some discrete random variable, n .

And in each box, there are some number of widgets. And again, this is also random. And in each box, say for Box I , there are x_i number of widgets in each one. What we're really interested in in this problem is, how many widgets are there total in this crate?

So in the crate, there are boxes, and in the boxes, there are widgets. How many widgets are there total within the crate? And we'll call that a random variable, t . And the problem gives us some information.

It tells us that the expectation of the number of widgets in each box for all the boxes is the same. It's 10. And also, the expectation of the number of boxes is also 10. And furthermore, the variance of x of the number of widgets and the number of boxes is all 16.

And lastly, an important fact is that all the x_i 's, so all the widgets for each box, and the total number of boxes, these random variables are all independent. So to calculate t , t is just a sum of x_1 through x_n . So x_1 is the number of widgets in Box 1, x_2 is the number of widgets in Box 2, and all the way through Box n .

So what makes this difficult is that the n is actually random. We don't actually know how many boxes there are. So we don't even know how many terms there are in the sum. Well, let's take a slightly simpler problem.

Let's pretend that we actually know there are exactly 12 boxes. And in that case, the only thing that's random now is how many widgets there are in each box. And so let's call [? sum ?] a new random variable, s , the sum of x_1 through x_{12} .

So this would tell us, this is the number of widgets in 12 boxes. All right. And because each of these x_i 's are independent, and they have the same expectation, just by linearity of expectations, we know that the expectation of s is just 12 copies of the same expectation of x_i . And similarly, because we also assume that all the x_i 's are independent, the variance of s , we can just add the variances of each of these terms.

So again, there are 12 copies of the variance of x_i . So we've done a simpler version of this problem, where we've assumed we know what n is, that n is 12. And we've seen that in this simpler case, it's pretty simple to calculate what the expectation of the sum is.

So let's try to use that knowledge to help us calculate the actual problem, where n is actually random. So what we'll do is use the law of iterated expectations. And so this is written in terms of x and y , but we can very easily just substitute in for the random variables that we care about. Where in this case, what we see is that in order to build things up, it would be helpful if we condition on something that is useful.

And in this case, it's fairly clear that it would be helpful if we condition on n , the number of boxes. So if we knew how many boxes there were, then we can drop down to the level of widgets within each box. And then once we have that, we can build up and average over the total number of boxes.

So what we should do is condition on n , the number of boxes. So what have we discovered through this simpler exercise earlier? Well, we've discovered that if we knew the number of boxes, then the expectation of the total number of widgets is just the number of boxes times the number of widgets in each one, or the expectation of the number of widgets in each one.

So we can use that information to help us here. Because now, this is basically the same scenario, except that the number of boxes is now random. Instead of being 12, it could be anything. But if we just condition on the number of boxes being equal to n , then we know that there are exactly n copies of this.

But notice that n here is still random. And so what we get is that the expectation is n times the expectation of the number of widgets in each box, which we know is 10. So it's expectation of 10 times n or 10 times the expectation of n , which gives us 100. Because there are, on expectation, 10 boxes.

So this, again, makes intuitive sense. Because we know that on average, there are 10 boxes. And on average, each box has 10 widgets inside. And so on average, we expect that there will be 100 widgets. And the key thing here is that we actually relied on this independence.

So if the number of widgets in each box vary depending on-- or if the distribution of the number of widgets in each box vary depending on how many boxes there were, then we wouldn't be able to do it this simply. OK, so that gives us the answer to the first part, the expectation of the total number of widgets. Now let's do the second part, which is the variance.

The variance, we'll again use this idea of conditioning and splitting things up, and use the law of total variance. So the variance of t is going to be equal to the expectation of the conditional variance plus the variance of the conditional expectation. So what we have to do now is just to calculate what all of these pieces are.

So let's start with this thing here, the conditional variance. So what is the conditional variance? Well, again, let's go back to our simpler case. We know that if we knew what n is, then the variance would just be n times the variance of each x_i . So what does that tell us?

That tells us that, well, if we knew what n was, so condition on n , the variance would just be n times the variance of each x_i . So we've just taken this analogy and generalized it to the case where we don't actually know what n is. We just condition on n , and we still have a random variable.

So then from that, we know that the expectation now, to get this first term, take the expectation of this conditional variance, it's just the expectation of n and the variance of x_i , we're given that. That's equal to 16. So it's n times 16, which we know is 160, because the expectation of n , we also know, is 10.

All right, let's do this second term now. We need the variance of the conditional expectation of t given n . Well, what is the conditional expectation of t given n ? We've already kind of used that here. And again, it's using the fact that if we knew what n was, the expectation would just be n times the expectation of the number of widgets in each box.

So it would be n times the expectation of each x_i . Now, to get the second term, we just take the variance of this. So the variance is the variance of n times the expectation of each x_i . And the expectation of each x_i is 10. So it's n times 10.

And now remember, when you calculate variances, [? if you ?] have a constant term inside, when you pull it out, you have to square it. So you get 100 times the variance of n . And we know that the variance of n is also 16. So this gives us 1600.

All right. So now we've calculated both terms here. The first term is equal to 160. The second term is equal to 1600. So to get the final answer, all we have to do is add this up. So we get that the final answer is equal to 1760.

And this is not as obvious as the expectation, where you could have just kind of guessed that it was equal to 100. So again, this was just another example of using conditioning and the laws of total variance and iterated expectations in order to help you solve a problem. And in this case, you could kind of see that there is a hierarchy, where you start with widgets.

Widgets are contained in boxes, and then crates contain some number of boxes. And so it's easy to just condition and do it level by level. So you condition on the number of boxes. If you know what the number of boxes are, then you can easily calculate how many widgets there are, on average.

And then you average over the number of boxes to get the final answer. So I hope that was helpful. And we'll see you next time.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

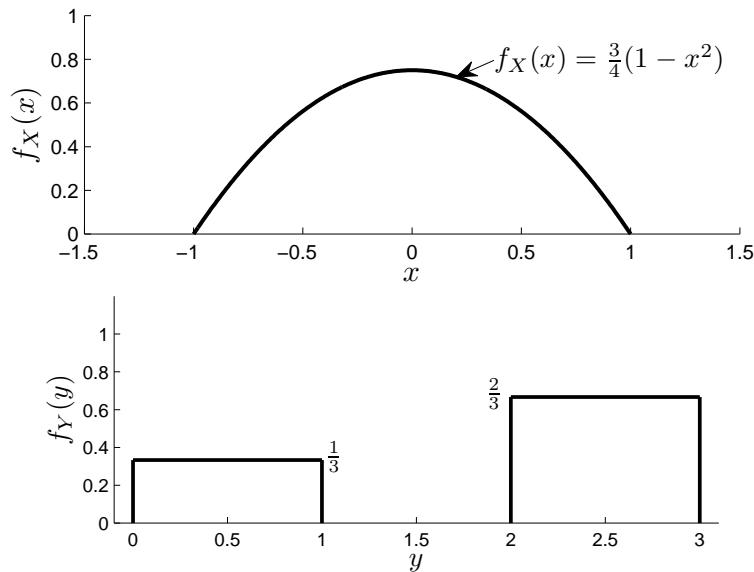
Problem Set 6
Due October 27, 2010

1. Random variables X and Y are distributed according to the joint PDF

$$f_{X,Y}(x,y) = \begin{cases} ax, & \text{if } 1 \leq x \leq 2 \text{ and } 0 \leq y \leq x, \\ 0, & \text{otherwise.} \end{cases}$$

- (a) Evaluate the constant a .
- (b) Determine the marginal PDF $f_Y(y)$.
- (c) Determine the conditional expectation of $1/X$ given that $Y = 3/2$.
- (d) Random variable Z is defined by $Z = Y - X$. Determine the PDF $f_Z(z)$.

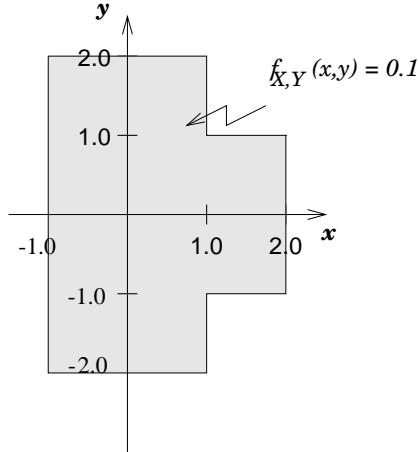
2. Let X and Y be two independent random variables. Their probability density functions are shown below.



Let $Z = X + Y$. Determine $f_Z(z)$.

3. Consider n independent tosses of a k -sided fair die. Let X_i be the number of tosses that result in i .
- (a) Are X_1 and X_2 uncorrelated, positively correlated, or negatively correlated? Give a one-line justification.
 - (b) Compute the covariance $\text{cov}(X_1, X_2)$ of X_1 and X_2 .

4. Random variables X and Y have the joint PDF shown below:



- (a) Find the conditional PDFs $f_{Y|X}(y | x)$ and $f_{X|Y}(x | y)$, for various values of x and y , respectively.
 (b) Find $\mathbf{E}[X | Y = y]$, $\mathbf{E}[X]$, and $\text{var}(X | Y = y)$. Use these to calculate $\text{var}(X)$.
 (c) Find $\mathbf{E}[Y | X = x]$, $\mathbf{E}[Y]$, and $\text{var}(Y | X = x)$. Use these to calculate $\text{var}(Y)$.
5. The wombat club has N members, where N is a random variable with PMF

$$p_N(n) = p^{n-1}(1-p) \quad \text{for } n = 1, 2, 3, \dots$$

On the second Tuesday night of every month, the club holds a meeting. Each wombat member attends the meeting with probability q , independently of all the other members. If a wombat attends the meeting, then it brings an amount of money, M , which is a continuous random variable with PDF

$$f_M(m) = \lambda e^{-\lambda m} \quad \text{for } m \geq 0.$$

N , M , and whether each wombat member attends are all independent. Determine:

- (a) The expectation and variance of the number of wombats showing up to the meeting.
 (b) The expectation and variance for the total amount of money brought to the meeting.

G1[†]. (a) Let $X_1, X_2, \dots, X_n, X_{n+1}, \dots, X_{2n}$ be independent and identically distributed random variables.

Find

$$\mathbf{E}[X_1 | X_1 + X_2 + \dots + X_n = x_0],$$

where x_0 is a constant.

- (b) Define

$$S_k = X_1 + X_2 + \dots + X_k, 1 \leq k \leq 2n.$$

Find

$$\mathbf{E}[X_1 | S_n = s_n, S_{n+1} = s_{n+1}, \dots, S_{2n} = s_{2n}],$$

where $s_n, s_{n+1}, \dots, s_{2n}$ are constants.

[†]Required for 6.431; optional for 6.041

MIT OpenCourseWare
<http://ocw.mit.edu>

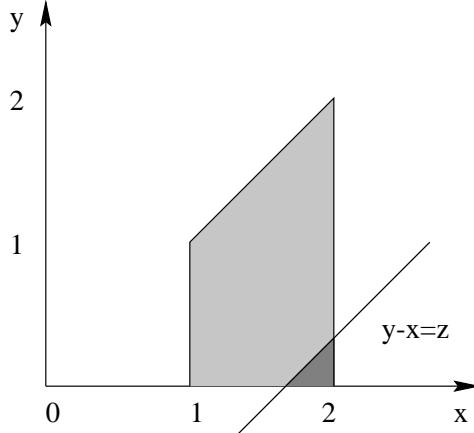
6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Problem Set 6: Solutions

1. Let us draw the region where $f_{X,Y}(x,y)$ is nonzero:



The joint PDF has to integrate to 1. From $\int_{x=1}^{x=2} \int_{y=0}^{y=x} ax dy dx = \frac{7}{3}a = 1$, we get $a = \frac{3}{7}$.

$$(b) f_Y(y) = \int f_{X,Y}(x,y) dy = \begin{cases} \int_1^2 \frac{3}{7}x dx, & \text{if } 0 \leq y \leq 1, \\ \int_y^2 \frac{3}{7}x dx, & \text{if } 1 < y \leq 2, \\ 0, & \text{otherwise} \end{cases} = \begin{cases} \frac{9}{14}, & \text{if } 0 \leq y \leq 1, \\ \frac{3}{14}(4-y^2), & \text{if } 1 < y \leq 2, \\ 0, & \text{otherwise.} \end{cases}$$

(c)

$$f_{X|Y}(x | \frac{3}{2}) = \frac{f_{X,Y}(x, \frac{3}{2})}{f_Y(\frac{3}{2})} = \frac{8}{7}x, \quad \text{for } \frac{3}{2} \leq x \leq 2 \text{ and 0 otherwise.}$$

Then,

$$\mathbf{E}\left[\frac{1}{X} | Y = \frac{3}{2}\right] = \int_{3/2}^2 \frac{1}{x} \frac{8}{7}x dx = \frac{4}{7}.$$

- (d) We use the technique of first finding the CDF and differentiating it to get the PDF.

$$\begin{aligned} F_Z(z) &= \mathbf{P}(Z \leq z) \\ &= \mathbf{P}(Y - X \leq z) \\ &= \begin{cases} 0, & \text{if } z < -2, \\ \int_{x=-z}^{x=2} \int_{y=0}^{y=x+z} \frac{3}{7}x dy dx = \frac{8}{7} + \frac{6}{7}z - \frac{1}{14}z^3, & \text{if } -2 \leq z \leq -1, \\ \int_{x=1}^{x=2} \int_{y=0}^{y=x+z} \frac{3}{7}x dy dx = 1 + \frac{9}{14}z, & \text{if } -1 < z \leq 0, \\ 1, & \text{if } 0 < z. \end{cases} \end{aligned}$$

$$f_Z(z) = \frac{d}{dz} F_Z(z) = \begin{cases} \frac{6}{7} - \frac{3}{14}z^2, & \text{if } -2 \leq z \leq -1, \\ \frac{9}{14}, & \text{if } -1 < z \leq 0, \\ 0, & \text{otherwise.} \end{cases}$$

2. The PDF of Z , $f_Z(z)$, can be readily computed using the convolution integral:

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(t)f_Y(z-t) dt.$$

For $z \in [-1, 0]$,

$$f_Z(z) = \int_{-1}^z \frac{1}{3} \cdot \frac{3}{4}(1-t^2) dt = \frac{1}{4} \left(z - \frac{z^3}{3} + \frac{2}{3} \right).$$

For $z \in [0, 1]$,

$$f_Z(z) = \int_{z-1}^z \frac{1}{3} \cdot \frac{3}{4}(1-t^2) dt = \frac{1}{4} \left(1 - \frac{z^3}{3} + \frac{(z-1)^3}{3} \right).$$

For $z \in [1, 2]$,

$$f_Z(z) = \int_{z-1}^1 \frac{1}{3} \cdot \frac{3}{4}(1-t^2) dt + \int_{-1}^{z-2} \frac{2}{3} \cdot \frac{3}{4}(1-t^2) dt = \frac{1}{4} \left(z + \frac{(z-1)^3}{3} - \frac{2(z-2)^3}{3} - 1 \right).$$

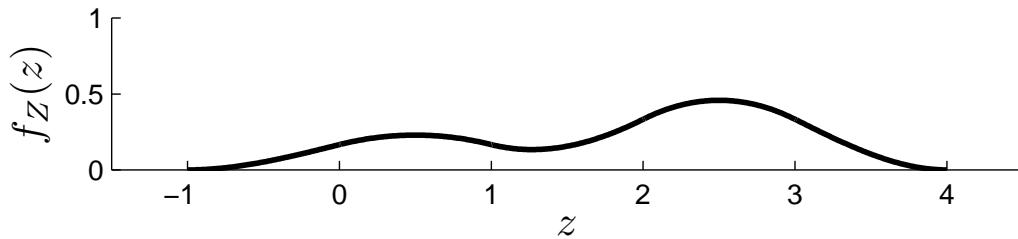
For $z \in [2, 3]$,

$$f_Z(z) = \int_{z-3}^{z-2} \frac{2}{3} \cdot \frac{3}{4}(1-t^2) dt = \frac{1}{6} (3 + (z-3)^3 - (z-2)^3).$$

For $z \in [3, 4]$,

$$f_Z(z) = \int_{z-3}^1 \frac{2}{3} \cdot \frac{3}{4}(1-t^2) dt = \frac{1}{6} (11 - 3z + (z-3)^3).$$

A sketch of $f_Z(z)$ is provided below.



3. (a) X_1 and X_2 are negatively correlated. Intuitively, a large number of tosses that result in a 1 suggests a smaller number of tosses that result in a 2.
 (b) Let A_t (respectively, B_t) be a Bernoulli random variable that is equal to 1 if and only if the t th toss resulted in 1 (respectively, 2). We have $\mathbf{E}[A_t B_s] = 0$ (since $A_t \neq 0$ implies $B_s = 0$) and

$$\mathbf{E}[A_t B_s] = \mathbf{E}[A_t] \mathbf{E}[B_s] = \frac{1}{k} \cdot \frac{1}{k} \quad \text{for } s \neq t.$$

Thus,

$$\begin{aligned} \mathbf{E}[X_1 X_2] &= \mathbf{E}[(A_1 + \dots + A_n)(B_1 + \dots + B_n)] \\ &= n \mathbf{E}[A_1(B_1 + \dots + B_n)] = n(n-1) \cdot \frac{1}{k} \cdot \frac{1}{k} \end{aligned}$$

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
 (Fall 2010)

and

$$\begin{aligned}\text{cov}(X_1, X_2) &= \mathbf{E}[X_1 X_2] - \mathbf{E}[X_1]\mathbf{E}[X_2] \\ &= \frac{n(n-1)}{k^2} - \frac{n^2}{k^2} = -\frac{n}{k^2}.\end{aligned}$$

The covariance of X_1 and X_2 is negative as expected.

4. (a) If X takes a value x between -1 and 1 , the conditional PDF of Y is uniform between -2 and 2 . If X takes a value x between 1 and 2 , the conditional PDF of Y is uniform between -1 and 1 .

Similarly, if Y takes a value y between -1 and 1 , the conditional PDF of X is uniform between -1 and 2 . If Y takes a value y between 1 and 2 , or between -2 and -1 , the conditional PDF of X is uniform between -1 and 1 .

- (b) We have

$$\mathbf{E}[X | Y = y] = \begin{cases} 0, & \text{if } -2 \leq y \leq -1, \\ 1/2, & \text{if } -1 < y \leq 1, \\ 0, & \text{if } 1 \leq y \leq 2, \end{cases}$$

and

$$\text{var}(X | Y = y) = \begin{cases} 1/3, & \text{if } -2 \leq y \leq -1, \\ 3/4, & \text{if } -1 < y \leq 1, \\ 1/3, & \text{if } 1 \leq y \leq 2. \end{cases}$$

It follows that $\mathbf{E}[X] = 3/10$ and $\text{var}(X) = 193/300$.

- (c) By symmetry, we have $\mathbf{E}[Y | X] = 0$ and $\mathbf{E}[Y] = 0$. Furthermore, $\text{var}(Y | X = x)$ is the variance of a uniform PDF (whose range depends on x), and

$$\text{var}(Y | X = x) = \begin{cases} 4/3, & \text{if } -1 \leq x \leq 1, \\ 1/3, & \text{if } 1 < x \leq 2. \end{cases}$$

Using the law of total variance, we obtain

$$\text{var}(Y) = \mathbf{E}[\text{var}(Y | X)] = \frac{4}{5} \cdot \frac{4}{3} + \frac{1}{5} \cdot \frac{1}{3} = 17/15.$$

5. First let us write out the properties of all of our random variables. Let us also define K to be the number of members attending a meeting and B to be the Bernoulli random variable describing whether or not a member attends a meeting.

$$\begin{aligned}\mathbf{E}[N] &= \frac{1}{1-p}, & \text{var}(N) &= \frac{p}{(1-p)^2}, \\ \mathbf{E}[M] &= \frac{1}{\lambda}, & \text{var}(M) &= \frac{1}{\lambda^2}, \\ \mathbf{E}[B] &= q, & \text{var}(B) &= q(1-q).\end{aligned}$$

- (a) Since $K = B_1 + B_2 + \dots + B_N$,

$$\begin{aligned}\mathbf{E}[K] &= \mathbf{E}[N] \cdot \mathbf{E}[B] = \frac{q}{1-p}, \\ \text{var}(K) &= \mathbf{E}[N] \cdot \text{var}(B) + (\mathbf{E}[B])^2 \cdot \text{var}(N) = \frac{q(1-q)}{1-p} + \frac{pq^2}{(1-p)^2}.\end{aligned}$$

(b) Let G be the total money brought to the meeting. Then $G = M_1 + M_2 + \dots + M_K$.

$$\begin{aligned}\mathbf{E}[G] &= \mathbf{E}[M] \cdot \mathbf{E}[K] = \frac{q}{\lambda(1-p)}, \\ \text{var}(G) &= \text{var}(M) \cdot \mathbf{E}[K] + (\mathbf{E}[M])^2 \text{var}(K) \\ &= \frac{q}{\lambda^2(1-p)} + \frac{1}{\lambda^2} \left(\frac{q(1-q)}{1-p} + \frac{pq^2}{(1-p)^2} \right).\end{aligned}$$

G1[†]. (a) Let X_1, X_2, \dots, X_n be independent, identically distributed (IID) random variables. We note that

$$\mathbf{E}[X_1 + \dots + X_n \mid X_1 + \dots + X_n = x_0] = x_0.$$

It follows from the linearity of expectations that

$$\begin{aligned}x_0 &= \mathbf{E}[X_1 + \dots + X_n \mid X_1 + \dots + X_n = x_0] \\ &= \mathbf{E}[X_1 \mid X_1 + \dots + X_n = x_0] + \dots + \mathbf{E}[X_n \mid X_1 + \dots + X_n = x_0]\end{aligned}$$

Because the X_i 's are identically distributed, we have the following relationship.

$$\mathbf{E}[X_i \mid X_1 + \dots + X_n = x_0] = \mathbf{E}[X_j \mid X_1 + \dots + X_n = x_0], \text{ for any } 1 \leq i \leq n, 1 \leq j \leq n.$$

Therefore,

$$\begin{aligned}n\mathbf{E}[X_1 \mid X_1 + \dots + X_n = x_0] &= x_0 \\ \mathbf{E}[X_1 \mid X_1 + \dots + X_n = x_0] &= \frac{x_0}{n}.\end{aligned}$$

(b) Note that we can rewrite $\mathbf{E}[X_1 \mid S_n = s_n, S_{n+1} = s_{n+1}, \dots, S_{2n} = s_{2n}]$ as follows:

$$\begin{aligned}&\mathbf{E}[X_1 \mid S_n = s_n, S_{n+1} = s_{n+1}, \dots, S_{2n} = s_{2n}] \\ &= \mathbf{E}[X_1 \mid S_n = s_n, X_{n+1} = s_{n+1} - s_n, X_{n+2} = s_{n+2} - s_{n+1}, \dots, X_{2n} = s_{2n} - s_{2n-1}] \\ &= \mathbf{E}[X_1 \mid S_n = s_n],\end{aligned}$$

where the last equality holds due to the fact that the X_i 's are independent. We also note that

$$\mathbf{E}[X_1 + \dots + X_n \mid S_n = s_n] = \mathbf{E}[S_n \mid S_n = s_n] = s_n.$$

It follows from the linearity of expectations that

$$\mathbf{E}[X_1 + \dots + X_n \mid S_n = s_n] = \mathbf{E}[X_1 \mid S_n = s_n] + \dots + \mathbf{E}[X_n \mid S_n = s_n].$$

Because the X_i 's are identically distributed, we have the following relationship:

$$\mathbf{E}[X_i \mid S_n = s_n] = \mathbf{E}[X_j \mid S_n = s_n], \text{ for any } 1 \leq i \leq n, 1 \leq j \leq n.$$

Therefore,

$$\mathbf{E}[X_1 + \dots + X_n \mid S_n = s_n] = n\mathbf{E}[X_1 \mid S_n = s_n] = s_n \Rightarrow \mathbf{E}[X_1 \mid S_n = s_n] = \frac{s_n}{n}.$$

[†]Required for 6.431; optional for 6.041

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

LECTURE 12

- **Readings:** Section 4.3;
parts of Section 4.5
(mean and variance only; no transforms)

Lecture outline

- Conditional expectation
 - Law of iterated expectations
 - Law of total variance
- Sum of a random number of independent r.v.'s
 - mean, variance

Conditional expectations

- Given the value y of a r.v. Y :
- $$E[X | Y = y] = \sum_x x p_{X|Y}(x | y)$$
- (integral in continuous case)
- Stick example: stick of length ℓ
break at uniformly chosen point Y
break again at uniformly chosen point X
 - $E[X | Y = y] = \frac{y}{2}$ (number)

$$E[X | Y] = \frac{Y}{2} \text{ (r.v.)}$$

- **Law of iterated expectations:**

$$E[E[X | Y]] = \sum_y E[X | Y = y] p_Y(y) = E[X]$$

- In stick example:
 $E[X] = E[E[X | Y]] = E[Y/2] = \ell/4$

var($X | Y$) and its expectation

- $\text{var}(X | Y = y) = E[(X - E[X | Y = y])^2 | Y = y]$
- $\text{var}(X | Y)$: a r.v.
with value $\text{var}(X | Y = y)$ when $Y = y$
- **Law of total variance:**

$$\text{var}(X) = E[\text{var}(X | Y)] + \text{var}(E[X | Y])$$

Proof:

- (a) Recall: $\text{var}(X) = E[X^2] - (E[X])^2$
- (b) $\text{var}(X | Y) = E[X^2 | Y] - (E[X | Y])^2$
- (c) $E[\text{var}(X | Y)] = E[X^2] - E[(E[X | Y])^2]$
- (d) $\text{var}(E[X | Y]) = E[(E[X | Y])^2] - (E[X])^2$

Sum of right-hand sides of (c), (d):

$$E[X^2] - (E[X])^2 = \text{var}(X)$$

Section means and variances

Two sections:

$y = 1$ (10 students); $y = 2$ (20 students)

$$y = 1 : \frac{1}{10} \sum_{i=1}^{10} x_i = 90 \quad y = 2 : \frac{1}{20} \sum_{i=11}^{30} x_i = 60$$

$$E[X] = \frac{1}{30} \sum_{i=1}^{30} x_i = \frac{90 \cdot 10 + 60 \cdot 20}{30} = 70$$

$$E[X | Y = 1] = 90, \quad E[X | Y = 2] = 60$$

$$E[X | Y] = \begin{cases} 90, & \text{w.p. } 1/3 \\ 60, & \text{w.p. } 2/3 \end{cases}$$

$$E[E[X | Y]] = \frac{1}{3} \cdot 90 + \frac{2}{3} \cdot 60 = 70 = E[X]$$

$$\begin{aligned} \text{var}(E[X | Y]) &= \frac{1}{3}(90 - 70)^2 + \frac{2}{3}(60 - 70)^2 \\ &= \frac{600}{3} = 200 \end{aligned}$$

Section means and variances (ctd.)

$$\frac{1}{10} \sum_{i=1}^{10} (x_i - 90)^2 = 10 \quad \frac{1}{20} \sum_{i=11}^{30} (x_i - 60)^2 = 20$$

$$\text{var}(X | Y = 1) = 10 \quad \text{var}(X | Y = 2) = 20$$

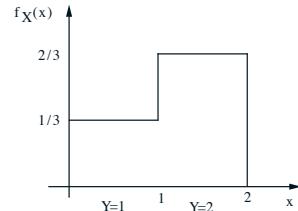
$$\text{var}(X | Y) = \begin{cases} 10, & \text{w.p. } 1/3 \\ 20, & \text{w.p. } 2/3 \end{cases}$$

$$\mathbb{E}[\text{var}(X | Y)] = \frac{1}{3} \cdot 10 + \frac{2}{3} \cdot 20 = \frac{50}{3}$$

$$\begin{aligned} \text{var}(X) &= \mathbb{E}[\text{var}(X | Y)] + \text{var}(\mathbb{E}[X | Y]) \\ &= \frac{50}{3} + 200 \\ &= (\text{average variability within sections}) \\ &\quad + (\text{variability between sections}) \end{aligned}$$

Example

$$\text{var}(X) = \mathbb{E}[\text{var}(X | Y)] + \text{var}(\mathbb{E}[X | Y])$$



$$\mathbb{E}[X | Y = 1] = \quad \quad \quad \mathbb{E}[X | Y = 2] =$$

$$\text{var}(X | Y = 1) = \quad \quad \quad \text{var}(X | Y = 2) =$$

$$\mathbb{E}[X] =$$

$$\text{var}(\mathbb{E}[X | Y]) =$$

Sum of a random number of independent r.v.'s

- N : number of stores visited (N is a nonnegative integer r.v.)

- X_i : money spent in store i

– X_i assumed i.i.d.

– independent of N

- Let $Y = X_1 + \dots + X_N$

$$\begin{aligned} \mathbb{E}[Y | N = n] &= \mathbb{E}[X_1 + X_2 + \dots + X_n | N = n] \\ &= \mathbb{E}[X_1 + X_2 + \dots + X_n] \\ &= \mathbb{E}[X_1] + \mathbb{E}[X_2] + \dots + \mathbb{E}[X_n] \\ &= n \mathbb{E}[X] \end{aligned}$$

- $\mathbb{E}[Y | N] = N \mathbb{E}[X]$

$$\begin{aligned} \mathbb{E}[Y] &= \mathbb{E}[\mathbb{E}[Y | N]] \\ &= \mathbb{E}[N \mathbb{E}[X]] \\ &= \mathbb{E}[N] \mathbb{E}[X] \end{aligned}$$

Variance of sum of a random number of independent r.v.'s

$$\bullet \text{ var}(Y) = \mathbb{E}[\text{var}(Y | N)] + \text{var}(\mathbb{E}[Y | N])$$

$$\bullet \mathbb{E}[Y | N] = N \mathbb{E}[X] \\ \text{var}(\mathbb{E}[Y | N]) = (\mathbb{E}[X])^2 \text{var}(N)$$

$$\bullet \text{ var}(Y | N = n) = n \text{var}(X) \\ \text{var}(Y | N) = N \text{var}(X) \\ \mathbb{E}[\text{var}(Y | N)] = \mathbb{E}[N] \text{var}(X)$$

$$\begin{aligned} \text{var}(Y) &= \mathbb{E}[\text{var}(Y | N)] + \text{var}(\mathbb{E}[Y | N]) \\ &= \mathbb{E}[N] \text{var}(X) + (\mathbb{E}[X])^2 \text{var}(N) \end{aligned}$$

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
 6.041SC Probabilistic Systems Analysis and Applied Probability
 Lecture 12 Bonus Video Solution

Problem 27.* We toss n times a biased coin whose probability of heads, denoted by q , is the value of a random variable Q with given mean μ and positive variance σ^2 . Let X_i be a Bernoulli random variable that models the outcome of the i th toss (i.e., $X_i = 1$ if the i th toss is a head). We assume that X_1, \dots, X_n are conditionally independent, given $Q = q$. Let X be the number of heads obtained in the n tosses.

- (a) Use the law of iterated expectations to find $\mathbf{E}[X_i]$ and $\mathbf{E}[X]$.
- (b) Find $\text{cov}(X_i, X_j)$. Are X_1, \dots, X_n independent?
- (c) Use the law of total variance to find $\text{var}(X)$. Verify your answer using the covariance result of part (b).

Solution. (a) We have, from the law of iterated expectations and the fact $\mathbf{E}[X_i | Q] = Q$,

$$\mathbf{E}[X_i] = \mathbf{E}[\mathbf{E}[X_i | Q]] = \mathbf{E}[Q] = \mu.$$

Since $X = X_1 + \dots + X_n$, it follows that

$$\mathbf{E}[X] = \mathbf{E}[X_1] + \dots + \mathbf{E}[X_n] = n\mu.$$

(b) We have, for $i = j$, using the conditional independence assumption,

$$/ \quad \mathbf{E}[X_i X_j | Q] = \mathbf{E}[X_i | Q] \mathbf{E}[X_j | Q] = Q^2,$$

and

$$\mathbf{E}[X_i X_j] = \mathbf{E}[\mathbf{E}[X_i X_j | Q]] = \mathbf{E}[Q^2].$$

Thus,

$$\text{cov}(X_i, X_j) = \mathbf{E}[X_i X_j] - \mathbf{E}[X_i] \mathbf{E}[X_j] = \mathbf{E}[Q^2] - \mu^2 = \sigma^2.$$

Since $\text{cov}(X_i, X_j) > 0$, X_1, \dots, X_n are not independent.

Also, for $i = j$, using the observation that $X_i^2 = X_i$,

$$\begin{aligned} \text{var}(X_i) &= \mathbf{E}[X_i^2] - (\mathbf{E}[X_i])^2 \\ &= \mathbf{E}[X_i] - (\mathbf{E}[X_i])^2 \\ &= \mu - \mu^2. \end{aligned}$$

(c) Using the law of total variance, and the conditional independence of X_1, \dots, X_n , we have

$$\begin{aligned}
 \text{var}(X) &= \mathbf{E} \text{ var}(X | Q) + \text{var } \mathbf{E}[X | Q] \\
 &= \mathbf{E} \text{ var}(X_1 + \dots + X_n | Q) + \text{var } \mathbf{E}[X_1 + \dots + X_n | Q] \\
 &= \mathbf{E} nQ(1 - Q) + \text{var}(nQ) \\
 &= n\mathbf{E}[Q - Q^2] + n^2 \text{var}(Q) \\
 &= n(\mu - \mu^2 - \sigma^2) + n^2 \sigma^2 \\
 &= n(\mu - \mu^2) + n(n - 1)\sigma^2.
 \end{aligned}$$

To verify the result using the covariance formulas of part (b), we write

$$\begin{aligned}
 \text{var}(X) &= \text{var}(X_1 + \dots + X_n) \\
 &= \sum_{i=1}^n \text{var}(X_i) + \sum_{\{(i,j) \mid i \neq j\}} \text{cov}(X_i, X_j) \\
 &= n\text{var}(X_1) + n(n - 1)\text{cov}(X_1, X_2) \\
 &= n(\mu - \mu^2) + n(n - 1)\sigma^2.
 \end{aligned}$$

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041SC Probabilistic Systems Analysis and Applied Probability
Lecture 12 Bonus Video Problem

Problem 27.* We toss n times a biased coin whose probability of heads, denoted by q , is the value of a random variable Q with given mean μ and positive variance σ^2 . Let X_i be a Bernoulli random variable that models the outcome of the i th toss (i.e., $X_i = 1$ if the i th toss is a head). We assume that X_1, \dots, X_n are conditionally independent, given $Q = q$. Let X be the number of heads obtained in the n tosses.

- (a) Use the law of iterated expectations to find $\mathbf{E}[X_i]$ and $\mathbf{E}[X]$.
- (b) Find $\text{cov}(X_i, X_j)$. Are X_1, \dots, X_n independent?
- (c) Use the law of total variance to find $\text{var}(X)$. Verify your answer using the covariance result of part (b).

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Recitation 13
October 21, 2010

For the problems below, recall the Law of Iterated Expectations and the Law of Total Variance:

$$\mathbf{E}[X] = \mathbf{E}[\mathbf{E}[X|Y]]$$

$$\text{var}(X) = \mathbf{E}[\text{var}(X|Y)] + \text{var}(\mathbf{E}[X|Y]).$$

1. Let X , Y , and Z be discrete random variables. Show the following generalizations of the law of iterated expectations.
 - (a) $\mathbf{E}[Z] = \mathbf{E}[\mathbf{E}[Z|X,Y]]$.
 - (b) $\mathbf{E}[Z|X] = \mathbf{E}[\mathbf{E}[Z|X,Y]|X]$.
 - (c) $\mathbf{E}[Z] = \mathbf{E}[\mathbf{E}[\mathbf{E}[Z|X,Y]|X]]$.
2. Example 4.17, page 223 in text.

We start with a stick of length ℓ . We break it at a point which is chosen randomly and uniformly over its length, and keep the piece that contains the left end of the stick. We then repeat the same process on the piece that we were left with.

- (a) What is the expected value of the length of the piece that we are left with after breaking twice?
- (b) What is the variance of the length of the piece that we are left with after breaking twice?
3. Widgets are stored in boxes, and then all boxes are assembled in a crate. Let X be the number of widgets in any particular box, and N be the number of boxes in a crate. Assume that X and N are independent integer-valued random variables, with expected value equal to 10, and variance equal to 16. Evaluate the expected value and variance of T , where T is the total number of widgets in a crate.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Recitation 13 Solutions
October 21, 2010

1. (a) We begin by writing the definition for $\mathbf{E}[Z | X, Y]$

$$\mathbf{E}[Z | X = x, Y = y] = \sum_z z p_{Z|X,Y}(z | x, y)$$

Since $\mathbf{E}[Z | X, Y]$ is a function of the random variables X and Y , and is equal to $\mathbf{E}[Z | X = x, Y = y]$ whenever $X = x$ and $Y = y$, which happens with probability $p_{X,Y}(x, y)$, using the expected value rule, we have

$$\begin{aligned} \mathbf{E}[\mathbf{E}[Z | X, Y]] &= \sum_x \sum_y \mathbf{E}[Z | X = x, Y = y] p_{X,Y}(x, y) \\ &= \sum_x \sum_y \sum_z z p_{Z|X,Y}(z | x, y) p_{X,Y}(x, y) \\ &= \sum_x \sum_y \sum_z z p_{X,Y,Z}(x, y, z) \\ &= \mathbf{E}[Z] \end{aligned}$$

- (b) We start with the definition for $\mathbf{E}[Z | X, Y]$ which is a function of the random variables X and Y , and is equal to $\mathbf{E}[Z | X = x, Y = y]$ whenever $X = x$ and $Y = y$, so

$$\mathbf{E}[Z | X = x, Y = y] = \sum_z z p_{Z|X,Y}(z | x, y)$$

Proceeding as above, but conditioning on the event $X = x$, we have

$$\begin{aligned} \mathbf{E}[\mathbf{E}[Z | X, Y = y] | X = x] &= \sum_y \mathbf{E}[Z | X = x, Y = y] p_{Y|X}(y | x) \\ &= \sum_y \sum_z z p_{Z|X,Y}(z | x, y) p_{Y|X}(y | x) \\ &= \sum_y \sum_z z p_{Y,Z|X}(y, z | x) \\ &= \mathbf{E}[Z | X = x] \end{aligned}$$

Since this is true for all possible values of x , we have $\mathbf{E}[\mathbf{E}[Z | Y, X] | X] = \mathbf{E}[Z | X]$.

- (c) We take expectations of both sides of the formula in part (b) to obtain

$$\mathbf{E}[\mathbf{E}[Z | X]] = \mathbf{E}[\mathbf{E}[\mathbf{E}[Z | X, Y] | X]].$$

By the law of iterated expectations, the left-hand side above is $\mathbf{E}[Z]$, which establishes the desired result.

2. Let Y be the length of the piece after we break for the first time. Let X be the length after we break for the second time.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
 (Fall 2010)

(a) The law of iterated expectations states:

$$\mathbf{E}[X] = \mathbf{E}[\mathbf{E}[X|Y]]$$

We have $\mathbf{E}[X|Y] = \frac{Y}{2}$ and $E[Y] = \frac{l}{2}$. So then:

$$\mathbf{E}[X] = \mathbf{E}[\mathbf{E}[X|Y]] = \mathbf{E}[Y/2] = \frac{1}{2}\mathbf{E}[Y] = \frac{1}{2}\frac{l}{2} = \frac{l}{4}$$

(b) We use the Law of Total Variance to find $\text{var}(X)$:

$$\text{var}(X) = \mathbf{E}[\text{var}(X | Y)] + \text{var}(\mathbf{E}[X | Y]).$$

Recall that the variance of a uniform random variable distributed over $[a, b]$ is $(b - a)^2/12$. Since Y is uniformly distributed over $[0, \ell]$, we have

$$\begin{aligned}\text{var}(Y) &= \frac{\ell^2}{12}, \\ \text{var}(X | Y) &= \frac{Y^2}{12}.\end{aligned}$$

We know that $\mathbf{E}[X | Y] = Y/2$, and so

$$\text{var}(\mathbf{E}[X | Y]) = \text{var}(Y/2) = \frac{1}{4}\text{var}(Y) = \frac{\ell^2}{48}.$$

Also,

$$\begin{aligned}\mathbf{E}[\text{var}(X | Y)] &= \mathbf{E}\left[\frac{Y^2}{12}\right] \\ &= \int_0^\ell \frac{y^2}{12} f_Y(y) dy \\ &= \frac{1}{12} \cdot \frac{1}{\ell} \int_0^\ell y^2 dy \\ &= \frac{\ell^2}{36}.\end{aligned}$$

Combining these results, we obtain

$$\text{var}(X) = \mathbf{E}[\text{var}(X | Y)] + \text{var}(\mathbf{E}[X | Y]) = \frac{\ell^2}{36} + \frac{\ell^2}{48} = \frac{7\ell^2}{144}.$$

3. Let X_i denote the number of widgets in the i^{th} box. Then $T = \sum_{i=1}^N X_i$.

$$\begin{aligned}\mathbf{E}[T] &= \mathbf{E}[\mathbf{E}[\sum_{i=1}^N X_i | N]] \\ &= \mathbf{E}[\sum_{i=1}^N \mathbf{E}[X_i | N]] \\ &= \mathbf{E}[\sum_{i=1}^N \mathbf{E}[X]] \\ &= \mathbf{E}[X] \cdot \mathbf{E}[N] = 100.\end{aligned}$$

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Fall 2010)

and,

$$\begin{aligned}\text{var}(T) &= \mathbf{E}[\text{var}(T|N)] + \text{var}(\mathbf{E}[T|N]) \\ &= \mathbf{E}\left[\text{var}\left(\sum_{i=1}^N X_i|N\right)\right] + \text{var}\left(\mathbf{E}\left[\sum_{i=1}^N X_i|N\right]\right) \\ &= \mathbf{E}[N\text{var}(X)] + \text{var}(N\mathbf{E}[X]) \\ &= (\text{var}(X))\mathbf{E}[N] + (\mathbf{E}[X])^2 \text{var}(N) \\ &= 16 \cdot 10 + 100 \cdot 16 = 1760.\end{aligned}$$

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Fall 2010)

Tutorial 6

October 21/22, 2010

1. Let X be a discrete random variable with PMF p_X and let Y be a continuous random variable, independent from X , with PDF f_Y . Derive a formula for the PDF of the random variable $X+Y$.
 2. The random variables X and Y are described by a joint PDF which is constant within the unit area quadrilateral with vertices $(0,0)$, $(0,1)$, $(1,2)$, and $(1,1)$. Use the law of total variance to find the variance of $X + Y$.
 3. (a) You roll a fair six-sided die, and then you flip a fair coin the number of times shown by the die. Find the expected value and the variance of the number of heads obtained.
(b) Repeat part (a) for the case where you roll two dice, instead of one.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Tutorial 6: Solutions

1. Let $Z = X + Y$. Using the 2 step CDF method,

$$\begin{aligned} F_Z(z) &= \mathbf{P}(Z \leq z) \\ &= \mathbf{P}(X + Y \leq z) \end{aligned}$$

Using the Total Probability Theorem, we have

$$\begin{aligned} F_Z(z) &= \sum_x p_X(x)p(x + Y \leq z) \\ &= \sum_x p_X(x)p(Y \leq z - x) \\ &= \sum_x p_X(x)F_Y(z - x) \end{aligned}$$

Differentiating both sides with respect to z , we obtain

$$\begin{aligned} f_Z(z) &= \frac{d}{dz}F_Z(z) \\ &= \sum_x p_X(x)f_Y(z - x) \end{aligned}$$

2. We will condition on X and use the law of total variance

$$\text{var}(X + Y) = \mathbf{E}[\text{var}(X + Y|X)] + \text{var}(\mathbf{E}[X + Y|X]).$$

Given a value x of X , the random variable Y is uniformly distributed in the interval $[x, x + 1]$, and the random variable $X + Y$ is uniformly distributed in the interval $[2x, 2x + 1]$. Therefore, $\mathbf{E}[X + Y|X] = 0.5 + 2X$ and $\text{var}(X + Y|X) = 1/12$. Thus,

$$\text{var}(X + Y) = \text{var}(0.5 + 2X) + \mathbf{E}[1/12] = 4\text{var}(X) + \mathbf{E}[1/12] = \frac{5}{12}.$$

3. (a) Let X_i be independent Bernoulli random variables that are equal to 1 if the i th flip results in heads. Let N be the number of coin flips. We have $\mathbf{E}[X_i] = 1/2$, $\text{var}(X_i) = 1/4$, $\mathbf{E}[N] = 7/2$, and $\text{var}(N) = 35/12$. (The last equality is obtained from the formula for the variance of a discrete uniform random variable.) Therefore, the expected number of heads is

$$\mathbf{E}[X_i]\mathbf{E}[N] = \frac{7}{4},$$

and the variance is

$$\text{var}(X_i)\mathbf{E}[N] + (\mathbf{E}[X_i])^2\text{var}(N) = \frac{1}{4} \cdot \frac{7}{2} + \frac{1}{4} \cdot \frac{35}{12} = \frac{77}{48}.$$

- (b) The experiment in part (b) can be viewed as consisting of two independent repetitions of the experiment in part (a). Thus, both the mean and the variance are doubled and become $7/2$ and $77/24$, respectively.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz 2 Solutions | Fall 2009)

(e) (7 points)

In the new universe in which $X = 2$, we are asked to compute the conditional PDF of Y given the event $Y \geq 3$.

$$f_{Y|X,Y \geq 3}(y | 2) = \frac{f_{Y|X}(y | 2)}{\mathbf{P}(Y \geq 3 | X = 2)}.$$

We first calculate the $\mathbf{P}(Y \geq 3 | X = 2)$.

$$\begin{aligned} \mathbf{P}(Y \geq 3 | X = 2) &= \int_3^{\infty} f_{Y|X}(y | 2) dy \\ &= \int_3^{\infty} 2e^{-2y} dy \\ &= 1 - F_{Y|X}(3 | 2) \\ &= 1 - (1 - e^{-2 \cdot 3}) \\ &= e^{-6}, \end{aligned}$$

where $F_{Y|X}(3 | 2)$ is the CDF of an exponential random variable with $\lambda = 2$ evaluated at $y = 3$. Substituting the values of $f_{Y|X}(y | 2)$ and $\mathbf{P}(Y \geq 3 | X = 2)$ yields

$$f_{Y|X,Y \geq 3}(y | 2) = \begin{cases} 2e^6 e^{-2y}, & y \geq 3 \\ 0, & \text{otherwise.} \end{cases}$$

Alternatively, $f_{Y|X}(y | 2)$ is an exponential random variable with $\lambda = 2$. To compute the conditional PMF $f_{Y|X,Y \geq 3}(y | 2)$, we can apply the memorylessness property of an exponential variable. Therefore, this conditional PMF is also an exponential random variable with $\lambda = 2$, but it is shifted by 3.

(f) (7 points)

Let's define $Z = e^{2X}$. Since X is an exponential random variable that takes on non-negative values ($X \geq 0$), $Z \geq 1$. We find the PDF of Z by first computing its CDF.

$$\begin{aligned} F_Z(z) &= \mathbf{P}(Z \leq z) \\ &= \mathbf{P}(e^{2X} \leq z) \\ &= \mathbf{P}(2X \leq \ln z) \\ &= \mathbf{P}\left(X \leq \frac{\ln z}{2}\right) \\ &= 1 - e^{-\frac{\ln z}{2}} \\ &= 1 - e^{\ln z^{-\frac{1}{2}}} \end{aligned}$$

The CDF of Z is:

$$F_Z(z) = \begin{cases} 1 - z^{-\frac{1}{2}} & z \geq 1 \\ 0, & z < 1 \end{cases}$$

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz 1 Solutions | Fall 2009)

Differentiating the CDF of Z yields the PDF

$$f_Z(z) = \begin{cases} \frac{1}{2}z^{-\frac{3}{2}} & z \geq 1 \\ 0, & z < 1 \end{cases}$$

Alternatively, you can apply the PDF formula for a strictly monotonic function of a continuous random variable. Recall if $z = g(x)$ and $x = h(z)$, then

$$f_Z(z) = f_X(h(z)) \left| \frac{dh}{dy}(z) \right|.$$

In this problem, $z = e^{2x}$ and $x = \frac{1}{2}\ln z$. Note that $f_Z(z)$ is nonzero for $z > 1$. Since X is an exponential random variable with $\lambda = 1$, $f_X(x) = e^x$. Thus,

$$\begin{aligned} f_Z(z) &= e^{-\frac{1}{2}\ln z} \left| \frac{1}{2z} \right| \\ &= e^{\ln z - \frac{1}{2}} \frac{1}{2z} \\ &= \frac{1}{2}z^{-\frac{3}{2}} \quad z \geq 1, \end{aligned}$$

where the second equality holds since the expression inside the absolute value is always positive for $z \geq 1$.

Problem 3. (10 points)

(a) (5 points) The quantity $\mathbf{E}[X | Y]$ is always:

- (i) A number.
- (ii) A discrete random variable.
- (iii) A continuous random variable.
- (iv) Not enough information to choose between (i)-(iii).

If X and Y are not independent, then $\mathbf{E}[X | Y]$ is a function of Y and is therefore a continuous random variable. However if X and Y are independent, then $\mathbf{E}[X | Y] = \mathbf{E}[X]$ which is a number.

(b) (5 points) The quantity $\mathbf{E}[\mathbf{E}[X | Y, N] | N]$ is always:

- (i) A number.
- (ii) A discrete random variable.
- (iii) A continuous random variable.
- (iv) Not enough information to choose between (i)-(iii).

If X , Y and N are not independent, then the inner expectation $G(Y, N) = \mathbf{E}[X | Y, N]$ is a function of Y and N . Furthermore $\mathbf{E}[G(Y, N) | N]$ is a function of N , a discrete random variable. If X , Y and N are independent, then the inner expectation $\mathbf{E}[X | Y, N] = \mathbf{E}[X]$, which is a number. The expectation of a number given N is still a number, which is a special case of a discrete random variable.

Problem 4. (25 points)

(a) (i) (5 points)

Using the Law of Iterated Expectations, we have

$$\mathbf{E}[X] = \mathbf{E}[\mathbf{E}[X \mid Q]] = \mathbf{E}[Q] = \frac{1}{2}.$$

(ii) (5 points)

X is a Bernoulli random variable with a mean $p = \frac{1}{2}$ and its variance is $\text{var}(X) = p(1 - p) = 1/4$.

(b) (7 points)

We know that $\text{cov}(X, Q) = \mathbf{E}[XQ] - \mathbf{E}[X]\mathbf{E}[Q]$, so first let's calculate $\mathbf{E}[XQ]$:

$$\mathbf{E}[XQ] = \mathbf{E}[\mathbf{E}[XQ \mid Q]] = \mathbf{E}[Q\mathbf{E}[X \mid Q]] = \mathbf{E}[Q^2] = \frac{1}{3}.$$

Therefore, we have

$$\text{cov}(X, Q) = \frac{1}{3} - \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{12}.$$

(c) (8 points)

Using Bayes' Rule, we have

$$f_{Q|X}(q \mid 1) = \frac{f_Q(q)p_{X|Q}(1 \mid q)}{p_X(1)} = \frac{f_Q(q)\mathbf{P}(X = 1 \mid Q = q)}{\mathbf{P}(X = 1)}, \quad 0 \leq q \leq 1.$$

Additionally, we know that

$$\mathbf{P}(X = 1 \mid Q = q) = q,$$

and that for Bernoulli random variables

$$\mathbf{P}(X = 1) = \mathbf{E}[X] = \frac{1}{2}.$$

Thus, the conditional PDF of Q given $X = 1$ is

$$\begin{aligned} f_{Q|X}(q \mid 1) &= \frac{1 \cdot q}{1/2} \\ &= \begin{cases} 2q, & 0 \leq q \leq 1, \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

Problem 5. (21 points)

(a) (7 points)

$$\begin{aligned} \mathbf{P}(S \geq 1) &= \mathbf{P}(\min\{X, Y\} \geq 1) = \mathbf{P}(X \geq 1 \text{ and } Y \geq 1) = \mathbf{P}(X \geq 1)\mathbf{P}(Y \geq 1) \\ &= (1 - F_X(1))(1 - F_Y(1)) = (1 - \Phi(1))^2 \approx (1 - 0.8413)^2 \approx 0.0252. \end{aligned}$$

(b) (7 points)

Recalling Problem 2 of Problem Set 6, we have

$$\begin{aligned}\mathbf{P}(s \leq S \text{ and } L \leq \ell) &= \mathbf{P}(s \leq \min\{X, Y\} \text{ and } \max\{X, Y\} \leq \ell) \\ &= \mathbf{P}(s \leq X \text{ and } s \leq Y \text{ and } X \leq \ell \text{ and } Y \leq \ell) \\ &= \mathbf{P}(s \leq X \leq \ell)\mathbf{P}(s \leq Y \leq \ell) \\ &= (F_X(\ell) - F_X(s))(F_Y(\ell)F_Y(s)).\end{aligned}$$

(c) (7 points)

Given that $s \leq s + \delta \leq \ell$, the event $\{s \leq S \leq s + \delta, \ell \leq L \leq \ell + \delta\}$ is made up of the union of two disjoint possible events:

$$\{s \leq X \leq s + \delta, \ell \leq Y \leq \ell + \delta\} \cup \{s \leq Y \leq s + \delta, \ell \leq X \leq \ell + \delta\}.$$

In other words, either $S = X$ and $L = Y$, or $S = Y$ and $L = X$. Because the two events are disjoint, the probability of their union is equal to the sum of their individual probabilities.

Using also the independence of X and Y , we have

$$\begin{aligned}\mathbf{P}(s \leq S \leq s + \delta, \ell \leq L \leq \ell + \delta) &= \mathbf{P}(s \leq X \leq s + \delta, \ell \leq Y \leq \ell + \delta) \\ &\quad + \mathbf{P}(s \leq Y \leq s + \delta, \ell \leq X \leq \ell + \delta) \\ &= \mathbf{P}(s \leq X \leq s + \delta)\mathbf{P}(\ell \leq Y \leq \ell + \delta) \\ &\quad + \mathbf{P}(s \leq Y \leq s + \delta)\mathbf{P}(\ell \leq X \leq \ell + \delta) \\ &= \int_s^{s+\delta} f_X(x)dx \int_\ell^{\ell+\delta} f_Y(y)dy \\ &\quad + \int_s^{s+\delta} f_Y(y)dy \int_\ell^{\ell+\delta} f_X(x)dx\end{aligned}$$

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Question 1

Multiple choice questions. **CLEARLY** circle the best answer for each question below. Each question is worth 4 points each, with no partial credit given.

- a. Let X_1 , X_2 , and X_3 be independent random variables with the continuous uniform distribution over $[0, 1]$. Then $\mathbf{P}(X_1 < X_2 < X_3) =$

- (i) 1/6
- (ii) 1/3
- (iii) 1/2
- (iv) 1/4

Solution: To understand the principle, first consider a simpler problem with X_1 and X_2 as given above. Note that $\mathbf{P}(X_1 < X_2) + \mathbf{P}(X_2 < X_1) + \mathbf{P}(X_1 = X_2) = 1$ since the corresponding events are disjoint and exhaust all the possibilities. But $\mathbf{P}(X_1 < X_2) = \mathbf{P}(X_2 < X_1)$ by symmetry. Furthermore, $\mathbf{P}(X_1 = X_2) = 0$ since the random variables are continuous. Therefore, $\mathbf{P}(X_1 < X_2) = 1/2$.

Analogously, omitting the events with zero probability but making sure to exhaust all other possibilities, we have that $\mathbf{P}(X_1 < X_2 < X_3) + \mathbf{P}(X_1 < X_3 < X_2) + \mathbf{P}(X_2 < X_1 < X_3) + \mathbf{P}(X_2 < X_3 < X_1) + \mathbf{P}(X_3 < X_1 < X_2) + \mathbf{P}(X_3 < X_2 < X_1) = 1$. And, by symmetry, $\mathbf{P}(X_1 < X_2 < X_3) = \mathbf{P}(X_1 < X_3 < X_2) = \mathbf{P}(X_2 < X_1 < X_3) = \mathbf{P}(X_2 < X_3 < X_1) = \mathbf{P}(X_3 < X_1 < X_2) = \mathbf{P}(X_3 < X_2 < X_1)$. Thus, $\mathbf{P}(X_1 < X_2 < X_3) = 1/6$.

- b. Let X and Y be two continuous random variables. Then

- (i) $\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y]$
- (ii) $\mathbf{E}[X^2 + Y^2] = \mathbf{E}[X^2] + \mathbf{E}[Y^2]$
- (iii) $f_{X+Y}(x+y) = f_X(x)f_Y(y)$
- (iv) $\text{var}(X+Y) = \text{var}(X) + \text{var}(Y)$

Solution: Since X^2 and Y^2 are random variables, the result follows by the linearity of expectation.

- c. Suppose X is uniformly distributed over $[0, 4]$ and Y is uniformly distributed over $[0, 1]$. Assume X and Y are independent. Let $Z = X + Y$. Then

- (i) $f_Z(4.5) = 0$
- (ii) $f_Z(4.5) = 1/8$
- (iii) $f_Z(4.5) = 1/4$
- (iv) $f_Z(4.5) = 1/2$

Solution: Since X and Y are independent, the result follows by convolution:

$$f_Z(4.5) = \int_{-\infty}^{\infty} f_X(\alpha) f_Y(4.5 - \alpha) d\alpha = \int_{3.5}^4 \frac{1}{4} d\alpha = \frac{1}{8}.$$

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz 2 Solutions | Spring 2008)

d. For the random variables defined in part (c), $\mathbf{P}(\max(X, Y) > 3)$ is equal to

- (i) 0
- (ii) 9/4
- (iii) 3/4
- (iv) 1/4

Solution: Note that $\mathbf{P}(\max(X, Y) > 3) = 1 - \mathbf{P}(\max(X, Y) \leq 3) = 1 - \mathbf{P}(\{X \leq 3\} \cap \{Y \leq 3\})$. But, X and Y are independent, so $\mathbf{P}(\{X \leq 3\} \cap \{Y \leq 3\}) = \mathbf{P}(X \leq 3)\mathbf{P}(Y \leq 3)$. Finally, computing the probabilities, we have $\mathbf{P}(X \leq 3) = 3/4$ and $\mathbf{P}(Y \leq 3) = 1$. Thus, $\mathbf{P}(\max(X, Y) > 3) = 1 - 3/4 = 1/4$.

e. Recall the hat problem from lecture: N people put their hats in a closet at the start of a party, where each hat is uniquely identified. At the end of the party each person randomly selects a hat from the closet. Suppose N is a Poisson random variable with parameter λ . If X is the number of people who pick their own hats, then $\mathbf{E}[X]$ is equal to

- (i) λ
- (ii) $1/\lambda^2$
- (iii) $1/\lambda$
- (iv) 1

Solution: Let $X = X_1 + \dots + X_N$ where each X_i is the indicator function such that $X_i = 1$ if the i^{th} person picks their own hat and $X_i = 0$ otherwise. By the linearity of the expectation, $\mathbf{E}[X \mid N = n] = \mathbf{E}[X_1 \mid N = n] + \dots + \mathbf{E}[X_N \mid N = n]$. But $\mathbf{E}[X_i \mid N = n] = 1/n$ for all $i = 1, \dots, n$. Thus, $\mathbf{E}[X \mid N = n] = n\mathbf{E}[X_i \mid N = n] = 1$. Finally, $\mathbf{E}[X] = \mathbf{E}[\mathbf{E}[X \mid N = n]] = 1$.

f. Suppose X and Y are Poisson random variables with parameters λ_1 and λ_2 respectively, where X and Y are independent. Define $W = X + Y$, then

- (i) W is Poisson with parameter $\min(\lambda_1, \lambda_2)$
- (ii) W is Poisson with parameter $\lambda_1 + \lambda_2$
- (iii) W may not be Poisson but has mean equal to $\min(\lambda_1, \lambda_2)$
- (iv) W may not be Poisson but has mean equal to $\lambda_1 + \lambda_2$

Solution: The quickest way to obtain the answer is through transforms: $M_X(s) = e^{\lambda_1(e^s - 1)}$ and $M_Y(s) = e^{\lambda_2(e^s - 1)}$. Since X and Y are independent, we have that $M_W(s) = e^{\lambda_1(e^s - 1)}e^{\lambda_2(e^s - 1)} = e^{(\lambda_1 + \lambda_2)(e^s - 1)}$, which equals the transform of a Poisson random variable with mean $\lambda_1 + \lambda_2$.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz 2 Solutions | Spring 2008)

g. Let X be a random variable whose transform is given by $M_X(s) = (0.4 + 0.6e^s)^{50}$. Then

- (i) $\mathbf{P}(X = 0) = \mathbf{P}(X = 50)$
- (ii) $\mathbf{P}(X = 51) > 0$
- (iii) $\boxed{\mathbf{P}(X = 0) = (0.4)^{50}}$
- (iv) $\mathbf{P}(X = 50) = 0.6$

Solution: Note that $M_X(s)$ is the transform of a binomial random variable X with $n = 50$ trials and the probability of success $p = 0.6$. Thus, $\mathbf{P}(X = 0) = 0.4^{50}$.

h. Let $X_i, i = 1, 2, \dots$ be independent random variables all distributed according to the PDF $f_X(x) = x/8$ for $0 \leq x \leq 4$. Let $S = \frac{1}{100} \sum_{i=1}^{100} X_i$. Then $\mathbf{P}(S > 3)$ is approximately equal to

- (i) $1 - \Phi(5)$
- (ii) $\Phi(5)$
- (iii) $\boxed{1 - \Phi\left(\frac{5}{\sqrt{2}}\right)}$
- (iv) $\Phi\left(\frac{5}{\sqrt{2}}\right)$

Solution: Let $S = \frac{1}{100} \sum_{i=1}^{100} Y_i$ where Y_i is the random variable given by $Y_i = X_1/100$. Since Y_i are *iid*, the distribution of S is approximately normal with mean $\mathbf{E}[S]$ and variance $\text{var}(S)$.

Thus, $\mathbf{P}(S > 3) = 1 - \mathbf{P}(S \leq 3) \approx 1 - \Phi\left(\frac{3 - \mathbf{E}(S)}{\sqrt{\text{var}(S)}}\right)$. Now,

$$\begin{aligned}\mathbf{E}[X_i] &= \int_0^4 x \frac{x}{8} dx = \frac{x^3}{24} \Big|_0^4 = \frac{8}{3} \\ \text{var}(X_i) &= \mathbf{E}[X_i^2] - (\mathbf{E}[X_i])^2 = \int_0^4 x^2 \frac{x}{8} dx - \left(\frac{8}{3}\right)^2 = \frac{x^4}{32} \Big|_0^4 - \left(\frac{8}{3}\right)^2 = \frac{8}{9}\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbf{E}[S] &= \frac{1}{100} \mathbf{E}[X_i] + \dots + \frac{1}{100} \mathbf{E}[X_{100}] = 8/3. \\ \text{var}(S) &= \frac{1}{100^2} \text{var}(X_i) + \dots + \frac{1}{100^2} \text{var}(X_i) = \frac{8}{9} \times \frac{1}{100}.\end{aligned}$$

and

$$\mathbf{P}(S > 3) \approx 1 - \Phi\left(\frac{3 - 8/3}{\sqrt{\frac{8}{9} \times \frac{1}{100}}}\right) = 1 - \Phi\left(\frac{5}{\sqrt{2}}\right).$$

i. Let $X_i, i = 1, 2, \dots$ be independent random variables all distributed according to the PDF $f_X(x) = 1, 0 \leq x \leq 1$. Define $Y_n = X_1 X_2 X_3 \dots X_n$, for some integer n . Then $\text{var}(Y_n)$ is equal to

- (i) $\frac{n}{12}$
- (ii) $\boxed{\frac{1}{3^n} - \frac{1}{4^n}}$
- (iii) $\frac{1}{12^n}$
- (iv) $\frac{1}{12}$

Solution: Since X_1, \dots, X_n are independent, we have that $\mathbf{E}[Y_n] = \mathbf{E}[X_1] \times \dots \times \mathbf{E}[X_n]$. Similarly, $\mathbf{E}[Y_n^2] = \mathbf{E}[X_1^2] \times \dots \times \mathbf{E}[X_n^2]$. Since $\mathbf{E}[X_i] = 1/2$ and $\mathbf{E}[X_i^2] = 1/3$ for $i = 1, \dots, n$, it follows that $\text{var}(S_n) = \mathbf{E}[Y_n^2] - (\mathbf{E}[Y_n])^2 = \frac{1}{3^n} - \frac{1}{4^n}$.

Question 2

Each Mac book has a lifetime that is exponentially distributed with parameter λ . The lifetime of Mac books are independent of each other. Suppose you have two Mac books, which you begin using at the same time. Define T_1 as the time of the first ~~laptop~~ failure and T_2 as the time of the second ~~laptop~~ failure.

- a. Compute $f_{T_1}(t_1)$.

Solution

Let M_1 be the life time of mac book 1 and M_2 the lifetime of mac book 2, where M_1 and M_2 are iid exponential random variables with CDF $F_M(m) = 1 - e^{-\lambda m}$. T_1 , the time of the first mac book failure, is the minimum of M_1 and M_2 . To derive the distribution of T_1 , we first find the CDF $F_{T_1}(t)$, and then differentiate to find the PDF $f_{T_1}(t)$.

$$\begin{aligned}
 F_{T_1}(t) &= P(\min(M_1, M_2) < t) \\
 &= 1 - P(\min(M_1, M_2) \geq t) \\
 &= 1 - P(M_1 \geq t)P(M_2 \geq t) \\
 &= 1 - (1 - F_M(t))^2 \\
 &= 1 - e^{-2\lambda t} \quad t \geq 0
 \end{aligned}$$

Differentiating $F_{T_1}(t)$ with respect to t yields:

$$f_{T_1}(t) = 2\lambda e^{-2\lambda t} \quad t \geq 0$$

- b. Let $X = T_2 - T_1$. Compute $f_{X|T_1}(x|t_1)$.

Solution

Conditioned on the time of the first mac book failure, the time until the other mac book fails is an exponential random variable by the memoryless property. The memoryless property tells us that regardless of the elapsed life time of the mac book, the time until failure has the same exponential CDF. Consequently,

$$f_{X|T_1}(x) = \lambda e^{-\lambda x} \quad x \geq 0.$$

- c. Is X independent of T_1 ? Give a mathematical justification for your answer.

Solution

Since we have shown in 2(c) that $f_{X|T_1}(x | t)$ does not depend on t , X and T_1 are independent.

- d. Compute $f_{T_2}(t_2)$ and $\mathbf{E}[T_2]$.

Solution

The time of the second laptop failure T_2 is equal to $T_1 + X$. Since X and T_1 were shown to be independent in 2(b), we convolve the densities found in 2(a) and 2(b) to determine $f_{T_2}(t)$.

$$\begin{aligned} f_{T_2}(t) &= \int_0^\infty f_{T_1}(\tau) f_X(t - \tau) d\tau \\ &= \int_0^t 2(\lambda)^2 e^{-2\lambda\tau} e^{-\lambda(t-\tau)} d\tau \\ &= 2\lambda e^{-\lambda t} \int_0^t \lambda e^{-\lambda\tau} d\tau \\ &= 2\lambda e^{-\lambda t} (1 - e^{-\lambda t}) \quad t \geq 0 \end{aligned}$$

Also, by the linearity of expectation, we have that $\mathbf{E}[T_2] = \mathbf{E}[T_1] + \mathbf{E}[X] = \frac{1}{2\lambda} + \frac{1}{\lambda} = \frac{3}{2\lambda}$.

An equivalent method for solving this problem is to note that T_2 is the maximum of M_1 and M_2 , and deriving the distribution of T_2 in our standard CDF to PDF method:

$$\begin{aligned} F_{T_2}(t) &= \mathbf{P}(\max(M_1, M_2) < t) \\ &= \mathbf{P}(M_2 \leq t) P(M_2 \leq t) \\ &= F_M(t)^2 \\ &= 1 - 2e^{-\lambda t} + e^{-2\lambda t} \quad t \geq 0 \end{aligned}$$

Differentiating $F_{T_2}(t)$ with respect to t yields:

$$f_{T_2}(t) = 2\lambda e^{-\lambda t} - 2\lambda e^{-2\lambda t} \quad t \geq 0$$

which is equivalent to our solution by convolution above.

Finally, from the above density we obtain that $\mathbf{E}[T_2] = \frac{2}{\lambda} - \frac{1}{2\lambda} = \frac{3}{2\lambda}$, which matches our earlier solution.

- e. Now suppose you have 100 Mac books, and let Y be the time of the first laptop failure. Find the best answer for $\mathbf{P}(Y < 0.01)$

Solution

Y is equal to the minimum of 100 independent exponential random variables. Following the derivation in (a), we determine by analogy:

$$f_Y(y) = 100\lambda e^{-100\lambda y} \quad t \geq 0$$

Integrating over y from 0 to .01, we find $\mathbf{P}(Y < .01) = 1 - e^{-\lambda}$.

Your friend, Charlie, loves Mac books so much he buys S new Mac books every day! On any given day S is equally likely to be 4 or 8, and all days are independent from each other. Let S_{100} be the number of Mac books Charlie buys over the next 100 days.

- f. (6 pts) Find the best approximation for $\mathbf{P}(S_{100} \leq 608)$. Express your final answer in terms of $\Phi(\cdot)$, the CDF of the standard normal.

Solution

Using the De Moivre - Laplace Approximation to the Binomial, and noting that the step size between values that S can take on is 4,

$$\begin{aligned} P(S_{100} \leq 608) &= \Phi\left(\frac{608 + 2 - 100 \times 6}{\sqrt{100 \times 4}}\right) \\ &= \Phi\left(\frac{10}{20}\right) \\ &= \Phi(.5) \end{aligned}$$

Question 3

Saif is a well intentioned though slightly indecisive fellow. Every morning he flips a coin to decide where to go. If the coin is heads he drives to the mall, if it comes up tails he volunteers at the local shelter. Saif's coin is not necessarily fair, rather it possesses a probability of heads equal to q . We do not know q , but we do know it is well-modeled by a random variable Q where the density of Q is

$$f_Q(q) = \begin{cases} 2q & \text{for } 0 \leq q \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Assume conditioned on Q each coin flip is independent. Note parts a, b, c, and $\{d, e\}$ may be answered independent of each other.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz 2 Solutions | Spring 2008)

a. (4 pts) What's the probability that Saif goes to the local shelter if he flips the coin once?

Solution

Let X_i be the outcome of a coin toss on the i^{th} trial, where $X_i = 1$ if the coin lands ‘heads’, and $X_i = 0$ if the coin lands ‘tails.’ By the total probability theorem:

$$\begin{aligned}\mathbf{P}(X_1 = 0) &= \int_0^1 P(X_1 = 0 \mid Q = q) f(q) dq \\ &= \int_0^1 (1 - q) 2q \, dq \\ &= \frac{1}{3}\end{aligned}$$

In an attempt to promote virtuous behavior, Saif’s father offers to pay him \$4 every day he volunteers at the local shelter. Define X as Saif’s payout if he flips the coin every morning for the next 30 days.

b. Find $\text{var}(X)$

Solution Let Y_i be a Bernoulli random variable describing the outcome of a coin tossed on morning i . Then, $Y_i = 1$ corresponds to the event that on morning i , Saif goes to the local shelter; $Y_i = 0$ corresponds to the event that on morning i , Saif goes to the mall. Assuming that the coin lands heads with probability q , i.e. that $Q = q$, we have that $P(Y_i = 1) = q$, and $P(Y_i = 0) = 1 - q$ for $i = 1, \dots, 30$.

Saif’s payout for next 30 days is described by random variable $X = 4(Y_1 + Y_2 + \dots + Y_{30})$.

$$\begin{aligned}\text{var}(X) &= 16 \text{ var}(Y_1 + Y_2 + \dots + Y_{30}) \\ &= 16 \text{ var}(\mathbf{E}[Y_1 + Y_2 + \dots + Y_{30} \mid Q]) + \mathbf{E}[\text{var}(Y_1 + Y_2 + \dots + Y_{30} \mid Q)]\end{aligned}$$

Now note that, conditioned on $Q = q$, Y_1, \dots, Y_{30} are independent. Thus, $\text{var}(Y_1 + Y_2 + \dots + Y_{30} \mid Q) = \text{var}(Y_1 \mid Q) + \dots + \text{var}(Y_{30} \mid Q)$. So,

$$\begin{aligned}\text{var}(X) &= 16 \text{ var}(30Q) + 16 \mathbf{E}[\text{var}(Y_1 \mid Q) + \dots + \text{var}(Y_{30} \mid Q)] \\ &= 16 \times 30^2 \text{var}(Q) + 16 \times 30 \mathbf{E}[Q(1 - Q)] \\ &= 16 \times 30^2 (\mathbf{E}[Q^2] - (\mathbf{E}[Q])^2) + 16 \times 30 (\mathbf{E}[Q] - \mathbf{E}[Q^2]) \\ &= 16 \times 30^2 (1/2 - 4/9) + 16 \times 30 (2/3 - 1/2) = 880\end{aligned}$$

since $\mathbf{E}[Q] = \int_0^1 2q \, dq = 2/3$ and $\mathbf{E}[Q^2] = \int_0^1 2q^2 \, dq = 1/2$.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz 2 Solutions | Spring 2008)

Let event B be that Saif goes to the local shelter at least once in k days.

- c. Find the conditional density of Q given B , $f_{Q|B}(q)$

Solution

By Bayes Rule:

$$\begin{aligned} f_{Q|B}(q) &= \frac{P(B | Q = q)f_Q(q)}{\int P(B | Q = q)f_Q(q)dq} \\ &= \frac{(1 - q^k)2q}{\int_0^1 (1 - q^k)2q dq} \\ &= \frac{2q(1 - q^k)}{1 - 2/(k + 2)} \quad 0 \leq q \leq 1 \end{aligned}$$

While shopping at the mall, Saif gets a call from his sister Mais. They agree to meet at the Coco Cabana Court yard at exactly 1:30PM. Unfortunately Mais arrives Z minutes late, where Z is a continuous uniform random variable from zero to 10 minutes. Saif is furious that Mais has kept him waiting, and demands Mais pay him R dollars, where $R = \exp(Z + 2)$.

- e. Find Saif's expected payout, $\mathbf{E}[R]$.

Solution

$$\begin{aligned} \mathbf{E}[R] &= \int_0^{10} e^{z+2}f(z)dz \\ &= \frac{e^2}{10} \int_0^{10} e^z dz \\ &= \frac{e^{12} - e^2}{10} \end{aligned}$$

- f. Find the density of Saif's payout, $f_R(r)$.

Solution

$$\begin{aligned} F_R(r) &= \mathbf{P}(e^{Z+2} \leq r) \\ &= \mathbf{P}(Z + 2 \leq \ln(r)) \\ &= \int_0^{\ln(r)-2} \frac{1}{10} dz \\ &= \frac{\ln(r) - 2}{10} \quad e^2 \leq r \leq e^{12} \end{aligned}$$

$$f_R(r) = \frac{1}{10r} \quad e^2 \leq r \leq e^{12}$$

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Problem 1. (80 points) In this problem:

- (i) X is a (continuous) uniform random variable on $[0, 4]$.
- (ii) Y is an exponential random variable, independent from X , with parameter $\lambda = 2$.

1. **(10 points)** Find the mean and variance of $X - 3Y$.
2. **(10 points)** Find the probability that $Y \geq X$.
(Let c be the answer to this question.)
3. **(10 points)** Find the conditional joint PDF of X and Y , given that the event $Y \geq X$ has occurred.
(You may express your answer in terms of the constant c from the previous part.)
4. **(10 points)** Find the PDF of $Z = X + Y$.
5. **(10 points)** Provide a fully labeled sketch of the conditional PDF of Z given that $Y = 3$.
6. **(10 points)** Find $\mathbf{E}[Z | Y = y]$ and $\mathbf{E}[Z | Y]$.
7. **(10 points)** Find the joint PDF $f_{Z,Y}$ of Z and Y .
8. **(10 points)** A random variable W is defined as follows. We toss a fair coin (independent of Y). If the result is “heads”, we let $W = Y$; if it is tails, we let $W = 2 + Y$. Find the probability of “heads” given that $W = 3$.

Problem 2. (30 points) Let X, X_1, X_2, \dots be independent normal random variables with mean 0 and variance 9. Let N be a positive integer random variable with $\mathbf{E}[N] = 2$ and $\mathbf{E}[N^2] = 5$. We assume that the random variables N, X, X_1, X_2, \dots are independent. Let $S = \sum_{i=1}^N X_i$.

1. **(10 points)** If δ is a small positive number, we have $\mathbf{P}(1 \leq |X| \leq 1 + \delta) \approx \alpha\delta$, for some constant α . Find the value of α .
2. **(10 points)** Find the variance of S .
3. **(5 points)** Are N and S uncorrelated? Justify your answer.
4. **(5 points)** Are N and S independent? Justify your answer.

Each question is repeated in the following pages. Please write your answer on the appropriate page.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz | Fall 2010)

Problem 1. (80 points) In this problem:

- (i) X is a (continuous) uniform random variable on $[0, 4]$.
- (ii) Y is an exponential random variable, independent from X , with parameter $\lambda = 2$.

1. **(10 points)** Find the mean and variance of $X - 3Y$.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz | Fall 2010)

2. **(10 points)** Find the probability that $Y \geq X$.

(Let c be the answer to this question.)

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz | Fall 2010)

3. **(10 points)** Find the conditional joint PDF of X and Y , given that the event $Y \geq X$ has occurred.

(You may express your answer in terms of the constant c from the previous part.)

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz | Fall 2010)

4. **(10 points)** Find the PDF of $Z = X + Y$.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz | Fall 2010)

5. **(10 points)** Provide a fully labeled sketch of the conditional PDF of Z given that $Y = 3$.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz | Fall 2010)

6. **(10 points)** Find $\mathbf{E}[Z | Y = y]$ and $\mathbf{E}[Z | Y]$.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz | Fall 2010)

7. **(10 points)** Find the joint PDF $f_{Z,Y}$ of Z and Y .

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz | Fall 2010)

8. **(10 points)** A random variable W is defined as follows. We toss a fair coin (independent of Y). If the result is “heads”, we let $W = Y$; if it is tails, we let $W = 2 + Y$. Find the probability of “heads” given that $W = 3$.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz | Fall 2010)

Problem 2. (30 points) Let X, X_1, X_2, \dots be independent normal random variables with mean 0 and variance 9. Let N be a positive integer random variable with $\mathbf{E}[N] = 2$ and $\mathbf{E}[N^2] = 5$. We assume that the random variables N, X, X_1, X_2, \dots are independent. Let $S = \sum_{i=1}^N X_i$.

1. **(10 points)** If δ is a small positive number, we have $\mathbf{P}(1 \leq |X| \leq 1 + \delta) \approx \alpha\delta$, for some constant α . Find the value of α .
 2. **(10 points)** Find the variance of S .

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz | Fall 2010)

3. **(5 points)** Are N and S uncorrelated? Justify your answer.

4. **(5 points)** Are N and S independent? Justify your answer.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
 (Quiz | Fall 2009)

	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767

The standard normal table. The entries in this table provide the numerical values of $\Phi(y) = \mathbf{P}(Y \leq y)$, where Y is a standard normal random variable, for y between 0 and 1.99. For example, to find $\Phi(1.71)$, we look at the row corresponding to 1.7 and the column corresponding to 0.01, so that $\Phi(1.71) = .9564$. When y is negative, the value of $\Phi(y)$ can be found using the formula $\Phi(y) = 1 - \Phi(-y)$.

Problem 2. (42 points)

The random variable X is exponential with parameter 1. Given the value x of X , the random variable Y is exponential with parameter equal to x (and mean $1/x$).

Note: Some useful integrals, for $\lambda > 0$:

$$\int_0^\infty xe^{-\lambda x} dx = \frac{1}{\lambda^2}, \quad \int_0^\infty x^2 e^{-\lambda x} dx = \frac{2}{\lambda^3}.$$

- (a) (7 points) Find the joint PDF of X and Y .
- (b) (7 points) Find the marginal PDF of Y .
- (c) (7 points) Find the conditional PDF of X , given that $Y = 2$.
- (d) (7 points) Find the conditional expectation of X , given that $Y = 2$.
- (e) (7 points) Find the conditional PDF of Y , given that $X = 2$ and $Y \geq 3$.
- (f) (7 points) Find the PDF of e^{2X} .

Problem 3. (10 points)

For the following questions, mark the correct answer. If you get it right, you receive 5 points for that question. You receive no credit if you get it wrong. A justification is not required and will not be taken into account.

Let X and Y be continuous random variables. Let N be a discrete random variable.

- (a) (5 points) The quantity $\mathbf{E}[X | Y]$ is always:
 - (i) A number.
 - (ii) A discrete random variable.
 - (iii) A continuous random variable.
 - (iv) Not enough information to choose between (i)-(iii).
- (b) (5 points) The quantity $\mathbf{E}[\mathbf{E}[X | Y, N] | N]$ is always:
 - (i) A number.
 - (ii) A discrete random variable.
 - (iii) A continuous random variable.
 - (iv) Not enough information to choose between (i)-(iii).

Problem 4. (25 points)

The probability of obtaining heads in a single flip of a certain coin is itself a random variable, denoted by Q , which is uniformly distributed in $[0, 1]$. Let $X = 1$ if the coin flip results in heads, and $X = 0$ if the coin flip results in tails.

- (a) (i) (5 points) Find the mean of X .
 (ii) (5 points) Find the variance of X .
- (b) (7 points) Find the covariance of X and Q .
- (c) (8 points) Find the conditional PDF of Q given that $X = 1$.

Problem 5. (21 points)

Let X and Y be **independent continuous** random variables with marginal PDFs f_X and f_Y , and marginal CDFs F_X and F_Y , respectively. Let

$$S = \min\{X, Y\}, \quad L = \max\{X, Y\}.$$

- (a) (7 points) If X and Y are standard normal, find the probability that $S \geq 1$.
- (b) (7 points) Fix some s and ℓ with $s \leq \ell$. Give a formula for

$$\mathbf{P}(s \leq S \text{ and } L \leq \ell)$$

involving F_X and F_Y , and no integrals.

- (c) (7 points) Assume that $s \leq s + \delta \leq \ell$. Give a formula for

$$\mathbf{P}(s \leq S \leq s + \delta, \ell \leq L \leq \ell + \delta),$$

as an integral involving f_X and f_Y .

Each question is repeated in the following pages. Please write your answer on the appropriate page.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz | Fall 2009)

Problem 2. (42 points)

The random variable X is exponential with parameter 1. Given the value x of X , the random variable Y is exponential with parameter equal to x (and mean $1/x$).

Note: Some useful integrals, for $\lambda > 0$:

$$\int_0^\infty xe^{-\lambda x} dx = \frac{1}{\lambda^2}, \quad \int_0^\infty x^2 e^{-\lambda x} dx = \frac{2}{\lambda^3}.$$

(a) (7 points) Find the joint PDF of X and Y .

(b) (7 points) Find the marginal PDF of Y .

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz | Fall 2009)

(c) (7 points) Find the conditional PDF of X , given that $Y = 2$.

(d) (7 points) Find the conditional expectation of X , given that $Y = 2$.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz | Fall 2009)

(e) (7 points) Find the conditional PDF of Y , given that $X = 2$ and $Y \geq 3$.

(f) (7 points) Find the PDF of e^{2X} .

Problem 3. (10 points)

For the following questions, mark the correct answer. If you get it right, you receive 5 points for that question. You receive no credit if you get it wrong. A justification is not required and will not be taken into account.

Let X and Y be continuous random variables. Let N be a discrete random variable.

(a) (5 points) The quantity $\mathbf{E}[X \mid Y]$ is always:

- (i) A number.
- (ii) A discrete random variable.
- (iii) A continuous random variable.
- (iv) Not enough information to choose between (i)-(iii).

(b) (5 points) The quantity $\mathbf{E}[\mathbf{E}[X \mid Y, N] \mid N]$ is always:

- (i) A number.
- (ii) A discrete random variable.
- (iii) A continuous random variable.
- (iv) Not enough information to choose between (i)-(iii).

Problem 4. (25 points)

The probability of obtaining heads in a single flip of a certain coin is itself a random variable, denoted by Q , which is uniformly distributed in $[0, 1]$. Let $X = 1$ if the coin flip results in heads, and $X = 0$ if the coin flip results in tails.

(a) (i) (5 points) Find the mean of X .

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz | Fall 2009)

(ii) (5 points) Find the variance of X .

(b) (7 points) Find the covariance of X and Q .

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz | Fall 2009)

(c) (8 points) Find the conditional PDF of Q given that $X = 1$.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz | Fall 2009)

Problem 5. (21 points)

Let X and Y be **independent continuous** random variables with marginal PDFs f_X and f_Y , and marginal CDFs F_X and F_Y , respectively. Let

$$S = \min\{X, Y\}, \quad L = \max\{X, Y\}.$$

- (a) (7 points) If X and Y are standard normal, find the probability that $S \geq 1$.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz | Fall 2009)

(b) (7 points) Fix some s and ℓ with $s \leq \ell$. Give a formula for

$$\mathbf{P}(s \leq S \text{ and } L \leq \ell)$$

involving F_X and F_Y , and no integrals.

(c) (7 points) Assume that $s \leq s + \delta \leq \ell$. Give a formula for

$$\mathbf{P}(s \leq S \leq s + \delta, \ell \leq L \leq \ell + \delta),$$

as an integral involving f_X and f_Y .

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041/6.431 Probabilistic Systems Analysis

Quiz II Review
Fall 2010

1

1 Probability Density Functions (PDF)

For a continuous RV X with PDF $f_X(x)$,

$$P(a \leq X \leq b) = \int_a^b f_X(x)dx$$

$$P(X \in A) = \int_A f_X(x)dx$$

Properties:

- Nonnegativity:

$$f_X(x) \geq 0 \quad \forall x$$

- Normalization:

$$\int_{-\infty}^{\infty} f_X(x)dx = 1$$

2

2 PDF Interpretation

Caution: $f_X(x) \neq P(X = x)$

- if X is continuous, $P(X = x) = 0 \quad \forall x!!$
- $f_X(x)$ can be ≥ 1

Interpretation: “probability per unit length” for “small” lengths around x

$$P(x \leq X \leq x + \delta) \approx f_X(x)\delta$$

3

3 Mean and variance of a continuous RV

$$E[X] = \int_{-\infty}^{\infty} xf_X(x)dx$$

$$\begin{aligned}\text{Var}(X) &= E[(X - E[X])^2] \\ &= \int_{-\infty}^{\infty} (x - E[X])^2 f_X(x)dx \\ &= E[X^2] - (E[X])^2 \quad (\geq 0)\end{aligned}$$

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f_X(x)dx$$

$$E[aX + b] = aE[X] + b$$

$$\text{Var}(aX + b) = a^2\text{Var}(X)$$

4

4 Cumulative Distribution Functions

Definition:

$$F_X(x) = P(X \leq x)$$

monotonically increasing from 0 (at $-\infty$) to 1 (at $+\infty$).

- Continuous RV (CDF is continuous in x):

$$F_X(x) = P(X \leq x) = \int_{-\infty}^x f_X(t)dt$$

$$f_X(x) = \frac{dF_X}{dx}(x)$$

- Discrete RV (CDF is piecewise constant):

$$F_X(x) = P(X \leq x) = \sum_{k \leq x} p_X(k)$$

$$p_X(k) = F_X(k) - F_X(k-1)$$

5

5 Uniform Random Variable

If X is a uniform random variable over the interval [a,b]:

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

$$F_X(x) = \begin{cases} 0 & \text{if } x \leq a \\ \frac{x-a}{b-a} & \text{if } a \leq x \leq b \\ 1 & \text{otherwise } (x > b) \end{cases}$$

$$E[X] = \frac{b-a}{2}$$

$$\text{var}(X) = \frac{(b-a)^2}{12}$$

6

6 Exponential Random Variable

X is an exponential random variable with parameter λ :

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

$$F_X(x) = \begin{cases} 1 - e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

$$E[X] = \frac{1}{\lambda} \quad \text{var}(X) = \frac{1}{\lambda^2}$$

Memoryless Property: Given that $X > t$, $X - t$ is an exponential RV with parameter λ

7

7 Normal/Gaussian Random Variables

General normal RV: $N(\mu, \sigma^2)$:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

$$E[X] = \mu, \quad \text{Var}(X) = \sigma^2$$

Property: If $X \sim N(\mu, \sigma^2)$ and $Y = aX + b$

$$\text{then } Y \sim N(a\mu + b, a^2\sigma^2)$$

8

8 Normal CDF

Standard Normal RV: $N(0, 1)$

CDF of standard normal RV Y at y: $\Phi(y)$

- given in tables for $y \geq 0$

- for $y < 0$, use the result: $\Phi(y) = 1 - \Phi(-y)$

To evaluate CDF of a general standard normal, express it as a function of a standard normal:

$$X \sim N(\mu, \sigma^2) \Leftrightarrow \frac{X - \mu}{\sigma} \sim N(0, 1)$$

$$P(X \leq x) = P\left(\frac{X - \mu}{\sigma} \leq \frac{x - \mu}{\sigma}\right) = \Phi\left(\frac{x - \mu}{\sigma}\right)$$

9

9 Joint PDF

Joint PDF of two continuous RV X and Y: $f_{X,Y}(x, y)$

$$P(A) = \int \int_A f_{X,Y}(x, y) dx dy$$

Marginal pdf: $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy$

$$E[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{X,Y}(x, y) dx dy$$

Joint CDF: $F_{X,Y}(x, y) = P(X \leq x, Y \leq y)$

10

10 Independence

By definition,

$$X, Y \text{ independent} \Leftrightarrow f_{X,Y}(x, y) = f_X(x)f_Y(y) \quad \forall (x, y)$$

If X and Y are independent:

- $E[XY] = E[X]E[Y]$
- $g(X)$ and $h(Y)$ are independent
- $E[g(X)h(Y)] = E[g(X)]E[h(Y)]$

11

11 Conditioning on an event

Let X be a continuous RV and A be an event with $P(A) > 0$,

$$f_{X|A}(x) = \begin{cases} \frac{f_X(x)}{P(X \in A)} & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases}$$

$$P(X \in B | X \in A) = \int_B f_{X|A}(x) dx$$

$$E[X|A] = \int_{-\infty}^{\infty} x f_{X|A}(x) dx$$

$$E[g(X)|A] = \int_{-\infty}^{\infty} g(x) f_{X|A}(x) dx$$

12

If A_1, \dots, A_n are disjoint events that form a partition of the sample space,

$$f_X(x) = \sum_{i=1}^n P(A_i) f_{X|A_i}(x) \text{ (}\approx \text{total probability theorem)}$$

$$E[X] = \sum_{i=1}^n P(A_i) E[X|A_i] \text{ (total expectation theorem)}$$

$$E[g(X)] = \sum_{i=1}^n P(A_i) E[g(X)|A_i]$$

13

12 Conditioning on a RV

X, Y continuous RV

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

$$f_X(x) = \int_{-\infty}^{\infty} f_Y(y) f_{X|Y}(x|y) dy \text{ (}\approx \text{total probthm)}$$

Conditional Expectation:

$$E[X|Y=y] = \int_{-\infty}^{\infty} x f_{X|Y}(x|y) dx$$

$$E[g(X)|Y=y] = \int_{-\infty}^{\infty} g(x) f_{X|Y}(x|y) dx$$

$$E[g(X, Y)|Y=y] = \int_{-\infty}^{\infty} g(x, y) f_{X|Y}(x|y) dx$$

14

Total Expectation Theorem:

$$E[X] = \int_{-\infty}^{\infty} E[X|Y=y] f_Y(y) dy$$

$$E[g(X)] = \int_{-\infty}^{\infty} E[g(X)|Y=y] f_Y(y) dy$$

$$E[g(X, Y)] = \int_{-\infty}^{\infty} E[g(X, Y)|Y=y] f_Y(y) dy$$

15

13 Continuous Bayes' Rule

X, Y continuous RV, N discrete RV, A an event.

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x) f_X(x)}{f_Y(y)} = \frac{f_{Y|X}(y|x) f_X(x)}{\int_{-\infty}^{\infty} f_{Y|X}(y|t) f_X(t) dt}$$

$$P(A|Y=y) = \frac{P(A) f_{Y|A}(y)}{f_Y(y)} = \frac{P(A) f_{Y|A}(y)}{f_{Y|A}(y) P(A) + f_{Y|A^c}(y) P(A^c)}$$

$$P(N=n|Y=y) = \frac{p_N(n) f_{Y|N}(y|n)}{f_Y(y)} = \frac{p_N(n) f_{Y|N}(y|n)}{\sum_i p_N(i) f_{Y|N}(y|i)}$$

16

14 Derived distributions

Def: PDF of a *function* of a RV X with known PDF: $Y = g(X)$.

Method:

- Get the CDF:

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = \int_{x|g(x) \leq y} f_X(x) dx$$

- Differentiate: $f_Y(y) = \frac{dF_Y}{dy}(y)$

Special case: if $Y = g(X) = aX + b$, $f_Y(y) = \frac{1}{|a|} f_X(\frac{y-b}{a})$

17

15 Convolution

$W = X + Y$, with X, Y independent.

- Discrete case:

$$p_W(w) = \sum_x p_X(x)p_Y(w-x)$$

- Continuous case:

$$f_W(w) = \int_{-\infty}^{\infty} f_X(x)f_Y(w-x) dx$$

18

Graphical Method:

- put the PMFs (or PDFs) on top of each other
- flip the PMF (or PDF) of Y
- shift the flipped PMF (or PDF) of Y by w
- cross-multiply and add (or evaluate the integral)

In particular, if X, Y are independent and normal, then
 $W = X + Y$ is normal.

19

16 Law of iterated expectations

$E[X|Y = y] = f(y)$ is a number.

$E[X|Y] = f(Y)$ is a random variable

(the expectation is taken with respect to X).

To compute $E[X|Y]$, first express $E[X|Y = y]$ as a function of y .

Law of iterated expectations:

$$E[X] = E[E[X|Y]]$$

(equality between two real numbers)

20

17 Law of Total Variance

$\text{Var}(X|Y)$ is a random variable that is a function of Y (the variance is taken with respect to X).

To compute $\text{Var}(X|Y)$, first express

$$\text{Var}(X|Y = y) = E[(X - E[X|Y = y])^2 | Y = y]$$

as a function of y .

Law of conditional variances:

$$\text{Var}(X) = E[\text{Var}(X|Y)] + \text{Var}(E[X|Y])$$

(equality between two real numbers)

21

18 Sum of a random number of iid RVs

N discrete RV, X_i i.i.d and independent of N .

$Y = X_1 + \dots + X_N$. Then:

$$E[Y] = E[X]E[N]$$

$$\text{Var}(Y) = E[N]\text{Var}(X) + (E[X])^2\text{Var}(N)$$

22

19 Covariance and Correlation

$$\begin{aligned}\text{Cov}(X, Y) &= E[(X - E[X])(Y - E[Y])] \\ &= E[XY] - E[X]E[Y]\end{aligned}$$

- By definition, X, Y are uncorrelated $\Leftrightarrow \text{Cov}(X, Y) = 0$.
- If X, Y independent $\Rightarrow X$ and Y are uncorrelated. (the converse is not true)
- In general, $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) + 2 \text{Cov}(X, Y)$
- If X and Y are uncorrelated, $\text{Cov}(X, Y) = 0$ and $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y)$

23

Correlation Coefficient: (dimensionless)

$$\rho = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} \in [-1, 1]$$

$\rho = 0 \Leftrightarrow X$ and Y are uncorrelated.

$|\rho| = 1 \Leftrightarrow X - E[X] = c[Y - E[Y]]$ (linearly related)

24

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
j o

Question 1: Multiple choice questions. **CLEARLY** circle the best answer for each question below. Each question is worth 4 points each, with no partial credit given.

a. (4 pts) Let X_1 , X_2 , and X_3 be independent random variables with the continuous uniform distribution over $[0, 1]$. Then $\mathbf{P}(X_1 < X_2 < X_3) =$

- (i) $1/6$
- (ii) $1/3$
- (iii) $1/2$
- (iv) $1/4$

b. (4 pts) Let X and Y be two continuous random variables. Then

- (i) $\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y]$
- (ii) $\mathbf{E}[X^2 + Y^2] = \mathbf{E}[X^2] + \mathbf{E}[Y^2]$
- (iii) $f_{X+Y}(x+y) = f_X(x)f_Y(y)$
- (iv) $\text{var}(X + Y) = \text{var}(X) + \text{var}(Y)$

c. (4 pts) Suppose X is uniformly distributed over $[0, 4]$ and Y is uniformly distributed over $[0, 1]$. Assume X and Y are independent. Let $Z = X + Y$. Then

- (i) $f_Z(4.5) = 0$
- (ii) $f_Z(4.5) = 1/8$
- (iii) $f_Z(4.5) = 1/4$
- (iv) $f_Z(4.5) = 1/2$

d. (4 pts) For the random variables defined in part (c), $\mathbf{P}(\max(X, Y) > 3)$ is equal to

- (i) 0
- (ii) $9/4$
- (iii) $3/4$
- (iv) $1/4$

e. (4 pts) Consider the following variant of the hat problem from lecture: N people put their hats in a closet at the start of a party, where each hat is uniquely identified. At the end of the party each person randomly selects a hat from the closet. Suppose N is a Poisson random variable with parameter λ . If X is the number of people who pick their own hats, then $\mathbf{E}[X]$ is equal to

- (i) λ
- (ii) $1/\lambda^2$
- (iii) $1/\lambda$
- (iv) 1

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
j o

f. (4 pts) Suppose X and Y are Poisson random variables with parameters λ_1 and λ_2 respectively, where X and Y are independent. Define $W = X + Y$, then

- (i) W is Poisson with parameter $\min(\lambda_1, \lambda_2)$
- (ii) W is Poisson with parameter $\lambda_1 + \lambda_2$
- (iii) W may not be Poisson but has mean equal to $\min(\lambda_1, \lambda_2)$
- (iv) W may not be Poisson but has mean equal to $\lambda_1 + \lambda_2$

g. (4 pts) Let X be a random variable whose transform is given by $M_X(s) = (0.4 + 0.6e^s)^{50}$. Then

- (i) $\mathbf{P}(X = 0) = \mathbf{P}(X = 50)$
- (ii) $\mathbf{P}(X = 51) > 0$
- (iii) $\mathbf{P}(X = 0) = (0.4)^{50}$
- (iv) $\mathbf{P}(X = 50) = 0.6$

h. (4 pts) Let $X_i, i = 1, 2, \dots$ be independent random variables all distributed according to the pdf $f_X(x) = x/8$ for $0 \leq x \leq 4$. Let $S = \frac{1}{100} \sum_{i=1}^{100} X_i$. Then $\mathbf{P}(S > 3)$ is approximately equal to

- (i) $1 - \Phi(5)$
- (ii) $\Phi(5)$
- (iii) $1 - \Phi\left(\frac{5}{\sqrt{2}}\right)$
- (iv) $\Phi\left(\frac{5}{\sqrt{2}}\right)$

i. (4 pts) Let $X_i, i = 1, 2, \dots$ be independent random variables all distributed according to the pdf $f_X(x) = 1, 0 \leq x \leq 1$. Define $Y_n = X_1 X_2 X_3 \dots X_n$, for some integer n . Then $\text{var}(Y_n)$ is equal to

- (i) $\frac{n}{12}$
- (ii) $\frac{1}{3^n} - \frac{1}{4^n}$
- (iii) $\frac{1}{12^n}$
- (iv) $\frac{1}{12}$

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
j o

Question 2: Each Mac book has a lifetime that is exponentially distributed with parameter λ . The lifetime of Mac books are independent of each other. Suppose you have two Mac books, which you begin using at the same time. Define T_1 as the time of the first ~~laptop~~ failure and T_2 as the time of the second ~~laptop~~ failure.

- a. (4 pts) Compute $f_{T_1}(t_1)$
- b. (5 pts) Let $X = T_2 - T_1$. Compute $f_{X|T_1}(x|t_1)$
- c. (5 pts) Is X independent of T_1 ? Give a mathematical justification for your answer.
- d. (8 pts) Compute $f_{T_2}(t_2)$ and $\mathbf{E}[T_2]$
- e. (5 pts) Now suppose you have 100 Mac books, and let Y be the time of the first laptop failure.
Find the best answer for $\mathbf{P}(Y < 0.01)$

Your friend, Charlie, loves Mac books so much he buys S new Mac books every day! On any given day S is equally likely to be 4 or 8, and all days are independent from each other. Let S_{100} be the number of Mac books Charlie buys over the next 100 days.

- f. (6 pts) Find the best approximation for $\mathbf{P}(S_{100} \leq 608)$. Express your final answer in terms of $\Phi(\cdot)$, the CDF of the standard normal.

Question 3: Saif is a well intentioned though slightly indecisive fellow. Every morning he flips a coin to decide where to go. If the coin is heads he drives to the mall, if it comes up tails he volunteers at the local shelter. Saif's coin is not necessarily fair, rather it possesses a probability of heads equal to q . We do not know q , but we do know it is well-modeled by a random variable Q where the density of Q is

$$f_Q(q) = \begin{cases} 2q & \text{for } 0 \leq q \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Assume conditioned on Q each coin flip is independent. Note parts a, b, c, and $\{d, e\}$ may be answered independent of each other.

- a. (4 pts) What's the probability that Saif goes to the local shelter if he flips the coin once?

In an attempt to promote virtuous behavior, Saif's father offers to pay him \$4 every day he volunteers at the local shelter. Define X as Saif's payout if he flips the coin every morning for the next 30 days.

- b. (6 pts) Find $\text{var}(X)$

Let event B be that Saif goes to the local shelter at least once in k days.

- c. (6 pts) Find the conditional density of Q given B , $f_{Q|B}(q)$

While shopping at the mall, Saif gets a call from his sister Mais. They agree to meet at the Coco Cabana Court yard at exactly 1:30PM. Unfortunately Mais arrives Z minutes late, where Z is a continuous uniform random variable from zero to 10 minutes. Saif is furious that Mais has kept him waiting, and demands Mais pay him R dollars where $R = \exp(Z + 2)$.

- d. (6 pts) Find Saif's expected payout, $\mathbf{E}[R]$

- e. (6 pts) Find the density of Saif's payout, $f_R(r)$

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz 2 Solutions | Fall 2010)

3. **(10 points)** Find the conditional joint PDF of X and Y , given that the event $Y \geq X$ has occurred.

(You may express your answer in terms of the constant c from the previous part.)

Let A be the event that $Y \geq X$. Since X and Y are independent,

$$\begin{aligned} f_{X,Y|A}(x,y) &= \frac{f_{X,Y}(x,y)}{\mathbf{P}(A)} = \frac{f_X(x)f_Y(y)}{\mathbf{P}(A)} \text{ for } (x,y) \in A \\ &= \begin{cases} \frac{4e^{-2y}}{1-e^{-8}}, & \text{if } 0 \leq x \leq 4, y \geq x \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

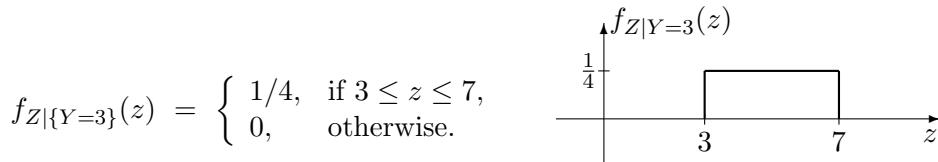
4. **(10 points)** Find the PDF of $Z = X + Y$.

Since X and Y are independent, the convolution integral can be used to find $f_Z(z)$.

$$\begin{aligned} f_Z(z) &= \int_{\max(0,z-4)}^z \frac{1}{4} 2e^{-2t} dt \\ &= \begin{cases} 1/4 \cdot (1 - e^{-2z}), & \text{if } 0 \leq z \leq 4, \\ 1/4 \cdot (e^8 - 1) e^{-2z}, & \text{if } z > 4, \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

5. **(10 points)** Provide a fully labeled sketch of the conditional PDF of Z given that $Y = 3$.

Given that $Y = 3$, $Z = X + 3$ and the conditional PDF of Z is a shifted version of the PDF of X . The conditional PDF of Z and its sketch are:



6. **(10 points)** Find $\mathbf{E}[Z | Y = y]$ and $\mathbf{E}[Z | Y]$.

The conditional PDF $f_{Z|Y=y}(z)$ is a uniform distribution between y and $y + 4$. Therefore,

$$\mathbf{E}[Z | Y = y] = y + 2.$$

The above expression holds true for all possible values of y , so

$$\mathbf{E}[Z | Y] = Y + 2.$$

7. **(10 points)** Find the joint PDF $f_{Z,Y}$ of Z and Y .

The joint PDF of Z and Y can be expressed as:

$$\begin{aligned} f_{Z,Y}(z,y) &= f_Y(y)f_{Z|Y}(z | y) \\ &= \begin{cases} 1/2 \cdot e^{-2y}, & \text{if } y \geq 0, y \leq z \leq y + 4, \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

8. **(10 points)** A random variable W is defined as follows. We toss a fair coin (independent of Y). If the result is “heads”, we let $W = Y$; if it is tails, we let $W = 2 + Y$. Find the probability of “heads” given that $W = 3$.

Let X be a Bernoulli random variable for the result of the fair coin where $X = 1$ if the coin lands “heads”. Because the coin is fair, $\mathbf{P}(X = 1) = \mathbf{P}(X = 0) = 1/2$. Furthermore, the conditional PDFs of W given the value of X are:

$$\begin{aligned} f_{W|X=1}(w) &= f_Y(w) \\ f_{W|X=0}(w) &= f_Y(w-2). \end{aligned}$$

Using the appropriate variation of Bayes’ Rule:

$$\begin{aligned} \mathbf{P}(X = 1 | W = 3) &= \frac{\mathbf{P}(X = 1)f_{W|X=1}(3)}{\mathbf{P}(X = 1)f_{W|X=1}(3) + \mathbf{P}(X = 0)f_{W|X=0}(3)} \\ &= \frac{\mathbf{P}(X = 1)f_Y(3)}{\mathbf{P}(X = 1)f_Y(3) + \mathbf{P}(X = 0)f_Y(1)} \\ &= \frac{\mathbf{P}(X = 1)f_Y(3)}{\mathbf{P}(X = 1)f_Y(3) + \mathbf{P}(X = 0)f_Y(1)} \\ &= \frac{e^{-6}}{e^{-6} + e^{-2}}. \end{aligned}$$

Problem 2. (30 points) Let X, X_1, X_2, \dots be independent normal random variables with mean 0 and variance 9. Let N be a positive integer random variable with $\mathbf{E}[N] = 2$ and $\mathbf{E}[N^2] = 5$. We assume that the random variables N, X, X_1, X_2, \dots are independent. Let $S = \sum_{i=1}^N X_i$.

1. **(10 points)** If δ is a small positive number, we have $\mathbf{P}(1 \leq |X| \leq 1 + \delta) \approx \alpha\delta$, for some constant α . Find the value of α .

$$\begin{aligned} \mathbf{P}(1 \leq |X| \leq 1 + \delta) &= 2\mathbf{P}(1 \leq X \leq 1 + \delta) \\ &\approx 2f_X(1)\delta. \end{aligned}$$

Therefore,

$$\begin{aligned} \alpha &= 2f_X(1) \\ &= 2 \cdot \frac{1}{\sqrt{9 \cdot 2\pi}} e^{-\frac{1}{2} \cdot \frac{(1-0)^2}{9}} \\ &= \frac{2}{3\sqrt{2\pi}} e^{-\frac{1}{18}}. \end{aligned}$$

2. **(10 points)** Find the variance of S .

Using the Law of Total Variance,

$$\begin{aligned} \text{var}(S) &= \mathbf{E}[\text{var}(S | N)] + \text{var}(\mathbf{E}[S | N]) \\ &= \mathbf{E}[9 \cdot N] + \text{var}(0 \cdot N) \\ &= 9\mathbf{E}[N] = 18. \end{aligned}$$

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Quiz 2 Solutions | Fall 2010)

3. **(5 points)** Are N and S uncorrelated? Justify your answer.

The covariance of S and N is

$$\begin{aligned}\text{cov}(S, N) &= \mathbf{E}[SN] - \mathbf{E}[S]\mathbf{E}[N] \\ &= \mathbf{E}[\mathbf{E}[SN | N]] - \mathbf{E}[\mathbf{E}[S | N]]\mathbf{E}[N] \\ &= \mathbf{E}[\mathbf{E}\left[\sum_{i=1}^N X_i N | N\right]] - \mathbf{E}[\mathbf{E}\left[\sum_{i=1}^N X_i | N\right]]\mathbf{E}[N] \\ &= \mathbf{E}[X_1]\mathbf{E}[N^2] - \mathbf{E}[X_1]\mathbf{E}[N] \\ &= 0\end{aligned}$$

since the $\mathbf{E}[X_1]$ is 0. Therefore, S and N are uncorrelated.

4. **(5 points)** Are N and S independent? Justify your answer.

S and N are not independent.

Proof: We have $\text{var}(S | N) = 9N$ and $\text{var}(S) = 18$, or, more generally, $f_{S|N}(s | n) = N(0, 9n)$ and $f_S(s) = N(0, 18)$ since a sum of independent normal random variables is also a normal random variable. Furthermore, since $\mathbf{E}[N^2] = 5 \neq (\mathbf{E}[N])^2 = 4$, N must take more than one value and is not simply a degenerate random variable equal to the number 2. In this case, N can take at least one value (with non-zero probability) that satisfies $\text{var}(S | N) = 9N \neq \text{var}(S) = 18$ and hence $f_{S|N}(s | n) \neq f_S(s)$. Therefore, S and N are not independent.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013 Transcript – Recitation: Bernoulli Process Practice

Hi everyone. Today I'm going to talk about Bernoulli process practice number one. In this problem, you are visiting a rain forest. But unfortunately you have run out of insect repellent.

As a result, the probability of you getting mosquito bites is really high. At each second, the probability that a mosquito will land on your neck is 0.5. If a mosquito lands on your neck, the probability that it will bite you is 0.2. And the probability that it will never bother you is 0.8. All of this happens independently among all mosquitoes.

For part A of the problem, we're interested in finding the expected value of the time between successive mosquito bites and the variance of the time between successive mosquito bites. From the problem statement we know that the probability distributions of getting mosquito bites at different times are identically distributed and independent. Therefore, the mosquito bites occur as a Bernoulli process with parameter p , where p represents the probability of getting a mosquito bite at each second.

And p can be calculated as the probability that a mosquito lands on your neck at each second multiplied by the probability that a mosquito will bite you, given that it has landed on your neck. And this is equal to 0.5 times 0.2, which is equal to 0.1.

Next let us define x as the time between successive mosquito bites. Because of the memory-less property of the Bernoulli process, which means the probability of getting mosquito bites at different times are independent, x is equivalent to the time until the next mosquito bite. And x is a geometrical random variable whose PMF is like the following.

For all x , let's say equal to 0, the probabilities are equal to 0. For x equal to 1, the probability that it takes 1 second to the next mosquito bite is simply equal to p . And for x equal to 2, the probability that it takes 2 seconds until the next mosquito bite is equal to 1 minus p times p . And for x equal to 3, the probability that it takes 3 seconds until the next mosquito bite is equal to 1 minus p to the power of 2 times p .

Similarly, for x equal to k , the probability that it takes k seconds until the next mosquito bite is equal to 1 minus p to the power of k minus 1 times p . Therefore the expected value of x is equal to 1 over p , which is equal to 1 over 0.1, which is equal to 10. And the variance of x is equal to 1 minus p over p squared, which is equal to 1 minus 0.1 over 0.1 squared, which is equal to 90.

For part B of the problem, we're considering another type of bug. Similar to the case as the mosquitoes, here at each second the probability that a tick will land on your neck is equal to 0.1. And if a tick lands on your neck, the probability that it will bite you is equal to 0.7. And the probability that it will never bother you is equal to 0.3. And all this happens independently among all ticks and all mosquitoes.

So similar to the case as part A, where mosquito bites occurs as a Bernoulli process with parameter p equal to 0.1, here the tick bites also across a Bernoulli process with parameter q equal to 0.1 times 0.7, which is equal to 0.07. And q is the probability of getting a tick bite at each second.

Therefore, the bug bites occurs as a merged process from the mosquito bites and the tick bites. And let r represent the parameter for the bug bites. So here r is equal to the probability of getting either a mosquito bite or a tick bite. And this is equivalent to 1 minus the probability of getting no mosquito bite and no tick bite.

Because the mosquito bites and the tick bites happens independently, therefore this can be written as 1 minus the probability of no mosquito bites times the probability of no tick bites at each second. And this is equal to 1 minus p times 1 minus q , which is p plus q minus pq , which is equal to 0.1 plus 0.7 minus 0.1 times 0.7. Which is equal to 0.163.

Next, let us define y as the time between successive bug bites. So similar as x in part a, here y is a geometric distribution with parameter r . And therefore the expected value of y is equal to 1 over r , which is equal to 1 over 0.163. That is approximately 6.135.

And the variance of y is equal to 1 minus r over r squared, which is equal to 1 minus 0.163 over 0.163 squared. And this is approximately 31.503. So this gives us the expected value of the time between successive bug bites and the variance of the time between successive bug bites. And this concludes our two days practice on Bernoulli process.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013

Transcript – Lecture 13

The following content is provided under a Creative Commons license. Your support will help MIT OpenCourseWare continue to offer high quality educational resources for free. To make a donation or view additional materials from hundreds of MIT courses, visit MIT OpenCourseWare at ocw.mit.edu.

PROFESSOR: So by now you have seen pretty much every possible trick there is in basic probability theory, about how to calculate distributions, and so on. You have the basic tools to do pretty much anything. So what's coming after this?

Well, probability is useful for developing the science of inference, and this is a subject to which we're going to come back at the end of the semester. Another chapter, which is what we will be doing over the next few weeks, is to deal with phenomena that evolve in time. So so-called random processes or stochastic processes. So what is this about?

So in the real world, you don't just throw two random variables and go home. Rather the world goes on. So you generate the random variable, then you get more random variables, and things evolve in time. And random processes are supposed to be models that capture the evolution of random phenomena over time. So that's what we will be doing.

Now when we have evolution in time, mathematically speaking, you can use discrete time or continuous time. Of course, discrete time is easier. And that's where we're going to start from. And we're going to start from the easiest, simplest random process, which is the so-called Bernoulli process, which is nothing but just a sequence of coin flips. You keep flipping a coin and keep going forever. That's what the Bernoulli process is.

So in some sense it's something that you have already seen. But we're going to introduce a few additional ideas here that will be useful and relevant as we go along and we move on to continuous time processes. So we're going to define the Bernoulli process, talk about some basic properties that the process has, and derive a few formulas, and exploit the special structure that it has to do a few quite interesting things.

By the way, where does the word Bernoulli come from? Well the Bernoulli's were a family of mathematicians, Swiss mathematicians and scientists around the 1700s. There were so many of them that actually-- and some of them had the same first name-- historians even have difficulty of figuring out who exactly did what. But in any case, you can imagine that at the dinner table they were probably flipping coins and doing Bernoulli trials. So maybe that was their pass-time.

OK. So what is the Bernoulli process? The Bernoulli process is nothing but a sequence of independent Bernoulli trials that you can think of as coin flips. So you can think the result of each trial being heads or tails. It's a little more convenient maybe to talk about successes and failures instead of heads or tails. Or if you wish numerical values, to use a 1 for a success and 0

for a failure. So the model is that each one of these trials has the same probability of success, p . And the other assumption is that these trials are statistically independent of each other.

So what could be some examples of Bernoulli trials? You buy a lottery ticket every week and you win or lose. Presumably, these are independent of each other. And if it's the same kind of lottery, the probability of winning should be the same during every week.

Maybe you want to model the financial markets. And a crude model could be that on any given day the Dow Jones is going to go up or down with a certain probability. Well that probability must be somewhere around 0.5, or so. This is a crude model of financial markets. You say, probably there is more into them. Life is not that simple. But actually it's a pretty reasonable model. It takes quite a bit of work to come up with more sophisticated models that can do better predictions than just pure heads and tails.

Now more interesting, perhaps to the examples we will be dealing with in this class-- a Bernoulli process is a good model for streams of arrivals of any kind to a facility. So it could be a bank, and you are sitting at the door of the bank. And at every second, you check whether a customer came in during that second or not. Or you can think about arrivals of jobs to a server. Or any other kind of requests to a service system.

So requests, or jobs, arrive at random times. You split the time into time slots. And during each time slot something comes or something does not come. And for many applications, it's a reasonable assumption to make that arrivals on any given slot are independent of arrivals in any other time slot. So each time slot can be viewed as a trial, where either something comes or doesn't come. And different trials are independent of each other.

Now there's two assumptions that we're making here. One is the independence assumption. The other is that this number, p , probability of success, is constant. Now if you think about the bank example, if you stand outside the bank at 9:30 in the morning, you'll see arrivals happening at a certain rate. If you stand outside the bank at 12:00 noon, probably arrivals are more frequent. Which means that the given time slot has a higher probability of seeing an arrival around noon time.

This means that the assumption of a constant p is probably not correct in that setting, if you're talking about the whole day. So the probability of successes or arrivals in the morning is going to be smaller than what it would be at noon. But if you're talking about a time period, let's say 10:00 to 10:15, probably all slots have the same probability of seeing an arrival and it's a good approximation. So we're going to stick with the assumption that p is constant, doesn't change with time.

Now that we have our model what do we do with it? Well, we start talking about the statistical properties that it has. And here there's two slightly different perspectives of thinking about what a random process is. The simplest version is to think about the random process as being just a sequence of random variables.

We know what random variables are. We know what multiple random variables are. So it's just an experiment that has associated with it a bunch of random variables. So once you have random variables, what do you do instinctively? You talk about the distribution of these random variables. We already specified for the Bernoulli process that each X_i is a Bernoulli random variable, with probability of success equal to p .

That specifies the distribution of the random variable X , or X_t , for general time t . Then you can calculate expected values and variances, and so on. So the expected value is, with probability p , you get a 1. And with probability $1 - p$, you get a 0. So the expected value is equal to p .

And then we have seen before a formula for the variance of the Bernoulli random variable, which is p times $1-p$. So this way we basically now have all the statistical properties of the random variable X_t , and we have those properties for every t . Is this enough of a probabilistic description of a random process? Well, no. You need to know how the different random variables relate to each other.

If you're talking about a general random process, you would like to know things. For example, the joint distribution of X_2 , with X_5 , and X_7 . For example, that might be something that you're interested in. And the way you specify it is by giving the joint PMF of these random variables. And you have to do that for every collection, or any subset, of the random variables you are interested in.

So to have a complete description of a random processes, you need to specify for me all the possible joint distributions. And once you have all the possible joint distributions, then you can answer, in principle, any questions you might be interested in.

How did we get around this issue for the Bernoulli process? I didn't give you the joint distributions explicitly. But I gave them to you implicitly. And this is because I told you that the different random variables are independent of each other. So at least for the Bernoulli process, where we make the independence assumption, we know that this is going to be the product of the PMFs. And since I have told you what the individual PMFs are, this means that you automatically know all the joint PMFs. And we can go to business based on that.

All right. So this is one view of what a random process is, just a collection of random variables. There's another view that's a little more abstract, which is the following. The entire process is to be thought of as one long experiment. So we go back to the chapter one view of probabilistic models. So there must be a sample space involved. What is the sample space?

If I do my infinite, long experiment of flipping an infinite number of coins, a typical outcome of the experiment would be a sequence of 0's and 1's. So this could be one possible outcome of the experiment, just an infinite sequence of 0's and 1's. My sample space is the set of all possible outcomes of this kind. Here's another possible outcome, and so on. And essentially we're dealing with a sample space, which is the space of all sequences of 0's and 1's. And we're making some sort of probabilistic assumption about what may happen in that experiment.

So one particular sequence that we may be interested in is the sequence of obtaining all 1's. So this is the sequence that gives you 1's forever. Once you take the point of view that this is our sample space-- its the space of all infinite sequences-- you can start asking questions that have to do with infinite sequences. Such as the question, what's the probability of obtaining the infinite sequence that consists of all 1's? So what is this probability? Let's see how we could calculate it.

So the probability of obtaining all 1's is certainly less than or equal to the probability of obtaining 1's, just in the first 10 tosses. OK. This is asking for more things to happen than this. If this event is true, then this is also true. Therefore the probability of this is smaller than the probability of that. This event is contained in that event. This implies this. So we have this inequality.

Now what's the probability of obtaining 1's in 10 trials? This is just p to the 10th because the trials are independent. Now of course there's no reason why I chose 10 here. The same argument goes through if I use an arbitrary number, k . And this has to be true for all k .

So this probability is less than p to the k , no matter what k I choose. Therefore, this must be less than or equal to the limit of this, as k goes to infinity. This is smaller than that for all k 's. Let k go to infinity, take k arbitrarily large, this number is going to become arbitrarily small. It goes to 0. And that proves that the probability of an infinite sequence of 1's is equal to 0.

So take limits of both sides. It's going to be less than or equal to the limit-- I shouldn't take a limit here. The probability is less than or equal to the limit of p to the k , as k goes to infinity, which is 0. So this proves in a formal way that the sequence of all 1's has 0 probability.

If you have an infinite number of coin flips, what's the probability that all of the coin flips result in heads? The probability of this happening is equal to zero. So this particular sequence has 0 probability. Of course, I'm assuming here that p is less than 1, strictly less than 1.

Now the interesting thing is that if you look at any other infinite sequence, and you try to calculate the probability of that infinite sequence, you would get a product of $(1-p)$ times 1, $1-p$ times 1, $1-p$, times p times p , times $1-p$ and so on. You keep multiplying numbers that are less than 1.

Again, I'm making the assumption that p is between 0 and 1. So $1-p$ is less than 1, p is less than 1. You keep multiplying numbers less than 1. If you multiply infinitely many such numbers, the infinite product becomes 0. So any individual sequence in this sample space actually has 0 probability. And that is a little bit counter-intuitive perhaps. But the situation is more like the situation where we deal with continuous random variables.

So if you could draw a continuous random variable, every possible outcome has 0 probability. And that's fine. But all of the outcomes collectively still have positive probability. So the situation here is very much similar. So the space of infinite sequences of 0's and 1's, that sample space is very much like a continuous space.

If you want to push that analogy further, you could think of this as the expansion of a real number. Or the representation of a real number in binary. Take a real number, write it down in

binary, you are going to get an infinite sequence of 0's and 1's. So you can think of each possible outcome here essentially as a real number.

So the experiment of doing an infinite number of coin flips is sort of similar to the experiment of picking a real number at random. When you pick real numbers at random, any particular real number has 0 probability. So similarly here, any particular infinite sequence has 0 probability.

So if we were to push that analogy further, there would be a few interesting things we could do. But we will not push it further. This is just to give you an indication that things can get pretty subtle and interesting once you start talking about random processes that involve forever, over the infinite time horizon. So things get interesting even in this context of the simple Bernoulli process.

Just to give you a preview of what's coming further, today we're going to talk just about the Bernoulli process. And you should make sure before the next lecture-- I guess between the exam and the next lecture-- to understand everything we do today. Because next time we're going to do everything once more, but in continuous time. And in continuous time, things become more subtle and a little more difficult. But we are going to build on what we understand for the discrete time case.

Now both the Bernoulli process and its continuous time analog has a property that we call memorylessness, whatever happened in the past does not affect the future. Later on in this class we're going to talk about more general random processes, so-called Markov chains, in which there are certain dependences across time. That is, what has happened in the past will have some bearing on what may happen in the future.

So it's like having coin flips where the outcome of the next coin flip has some dependence on the previous coin flip. And that gives us a richer class of models. And once we get there, essentially we will have covered all possible models.

So for random processes that are practically useful and which you can manipulate, Markov chains are a pretty general class of models. And almost any real world phenomenon that evolves in time can be approximately modeled using Markov chains. So even though this is a first class in probability, we will get pretty far in that direction.

All right. So now let's start doing a few calculations and answer some questions about the Bernoulli process. So again, the best way to think in terms of models that correspond to the Bernoulli process is in terms of arrivals of jobs to a facility. And there's two types of questions that you can ask. In a given amount of time, how many jobs arrived? Or conversely, for a given number of jobs, how much time did it take for them to arrive?

So we're going to deal with these two questions, starting with the first. For a given amount of time-- that is, for a given number of time periods-- how many arrivals have we had? How many of those X_i 's happen to be 1's?

We fix the number of time slots-- let's say n time slots-- and you measure the number of successes. Well this is a very familiar random variable. The number of successes in n independent coin flips-- or in n independent trials-- is a binomial random variable. So we know its distribution is given by the binomial PMF, and it's just this, for k going from 0 up to n .

And we know everything by now about this random variable. We know its expected value is n times p . And we know the variance, which is n times p , times $1-p$. So there's nothing new here. That's the easy part.

So now let's look at the opposite kind of question. Instead of fixing the time and asking how many arrivals, now let us fix the number of arrivals and ask how much time did it take. And let's start with the time until the first arrival.

So the process starts. We got our slots. And we see, perhaps, a sequence of 0's and then at some point we get a 1. The number of trials it took until we get a 1, we're going to call it T_1 . And it's the time of the first arrival. OK. What is the probability distribution of T_1 ? What kind of random variable is it?

We've gone through this before. The event that the first arrival happens at time little t is the event that the first $t-1$ trials were failures, and the trial number t happens to be a success. So for the first success to happen at time slot number 5, it means that the first 4 slots had failures and the 5th slot had a success.

So the probability of this happening is the probability of having failures in the first $t-1$ trials, and having a success at trial number 1. And this is the formula for t equal 1,2, and so on. So we know what this distribution is. It's the so-called geometric distribution.

Let me jump this through this for a minute. In the past, we did calculate the expected value of the geometric distribution, and it's $1/p$. Which means that if p is small, you expect to take a long time until the first success. And then there's a formula also for the variance of T_1 , which we never formally derived in class, but it was in your textbook and it just happens to be this.

All right. So nothing new until this point. Now, let's talk about this property, the memorylessness property. We kind of touched on this property when we discussed-- when we did the derivation in class of the expected value of T_1 .

Now what is the memoryless property? It's essentially a consequence of independence. If I tell you the results of my coin flips up to a certain time, this, because of independence, doesn't give you any information about the coin flips after that time. So knowing that we had lots of 0's here does not change what I believe about the future coin flips, because the future coin flips are going to be just independent coin flips with a given probability, p , for obtaining tails.

So this is a statement that I made about a specific time. That is, you do coin flips until 12 o'clock. And then at 12 o'clock, you start watching. No matter what happens before 12 o'clock, after 12:00, what you're going to see is just a sequence of independent Bernoulli trials with the same probability, p . Whatever happened in the past is irrelevant.

Now instead of talking about the fixed time at which you start watching, let's think about a situation where your sister sits in the next room, flips the coins until she observes the first success, and then calls you inside. And you start watching after this time. What are you're going to see?

Well, you're going to see a coin flip with probability p of success. You're going to see another trial that has probability p as a success, and these are all independent of each other. So what you're going to see starting at that time is going to be just a sequence of independent Bernoulli trials, as if the process was starting at this time. How long it took for the first success to occur doesn't have any bearing on what is going to happen afterwards. What happens afterwards is still a sequence of independent coin flips.

And this story is actually even more general. So your sister watches the coin flips and at some point tells you, oh, something really interesting is happening here. I got this string of a hundred 1's in a row. Come and watch. Now when you go in there and you start watching, do you expect to see something unusual?

There were unusual things that happened before you were called in. Does this means that you're going to see unusual things afterwards? No. Afterwards, what you're going to see is, again, just a sequence of independent coin flips. The fact that some strange things happened before doesn't have any bearing as to what is going to happen in the future.

So if the roulettes in the casino are properly made, the fact that there were 3 reds in a row doesn't affect the odds of whether in the next roll it's going to be a red or a black. So whatever happens in the past-- no matter how unusual it is-- at the time when you're called in, what's going to happen in the future is going to be just independent Bernoulli trials, with the same probability, p .

The only case where this story changes is if your sister has a little bit of foresight. So your sister can look ahead into the future and knows that the next 10 coin flips will be heads, and calls you before those 10 flips will happen. If she calls you in, then what are you going to see? You're not going to see independent Bernoulli trials, since she has psychic powers and she knows that the next ones would be 1's. She called you in and you will see a sequence of 1's. So it's no more independent Bernoulli trials.

So what's the subtle difference here? The future is independent from the past, provided that the time that you are called and asked to start watching is determined by someone who doesn't have any foresight, who cannot see the future. If you are called in, just on the basis of what has happened so far, then you don't have any information about the future.

And one special case is the picture here. You have your coin flips. Once you see a one that happens, once you see a success, you are called in. You are called in on the basis of what happened in the past, but without any foresight.

OK. And this subtle distinction is what's going to make our next example interesting and subtle. So here's the question. You buy a lottery ticket every day, so we have a Bernoulli process that's

running in time. And you're interested in the length of the first string of losing days. What does that mean?

So suppose that a typical sequence of events could be this one. So what are we discussing here? We're looking at the first string of losing days, where losing days means 0's. So the string of losing days is this string here. Let's call the length of that string, L . We're interested in the random variable, which is the length of this interval. What kind of random variable is it?

OK. Here's one possible way you might think about the problem. OK. Starting from this time, and looking until this time here, what are we looking at? We're looking at the time, starting from here, until the first success. So the past doesn't matter. Starting from here we have coin flips until the first success. The time until the first success in a Bernoulli process-- we just discussed that it's a geometric random variable.

So your first conjecture would be that this random variable here, which is 1 longer than the one we are interested in, that perhaps is a geometric random variable. And if this were so, then you could say that the random variable, L , is a geometric, minus 1. Can that be the correct answer?

A geometric random variable, what values does it take? It takes values 1, 2, 3, and so on. 1 minus a geometric would take values from 0, 1, 2, and so on. Can the random variable L be 0? No. The random variable L is the length of a string of losing days. So the shortest that L could be, would be just 1. If you get just one losing day and then you start winning, L would be equal to 1.

So L cannot be 0 by definition, which means that $L + 1$ cannot be 1, by definition. But if $L + 1$ were geometric, it could be equal to 1. Therefore this random variable, $L + 1$, is not a geometric.

OK. Why is it not geometric? I started watching at this time. From this time until the first success, that should be a geometric random variable. Where's the catch? If I'm asked to start watching at this time, it's because my sister knows that the next one was a failure. This is the time where the string of failures starts. In order to know that they should start watching here, it's the same as if I'm told that the next one is a failure.

So to be asked to start watching at this time requires that someone looked in the future. And in that case, it's no longer true that these will be independent Bernoulli trials. In fact, they're not. If you start watching here, you're certain that the next one is a failure. The next one is not an independent Bernoulli trial. That's why the argument that would claim that this $L + 1$ is geometric would be incorrect.

So if this is not the correct answer, which is the correct answer? The correct answer goes as follows. Your sister is watching. Your sister sees the first failure, and then tells you, OK, the failures-- or losing days-- have started. Come in and watch. So you start to watching at this time. And you start watching until the first success comes. This will be a geometric random variable.

So from here to here, this will be geometric. So things happen. You are asked to start watching. After you start watching, the future is just a sequence of independent Bernoulli trials. And the time until the first failure occurs, this is going to be a geometric random variable with parameter

p. And then you notice that the interval of interest is exactly the same as the length of this interval. This starts one time step later, and ends one time step later. So conclusion is that L is actually geometric, with parameter p.

OK, it looks like I'm missing one slide. Can I cheat a little from here?

OK. So now that we dealt with the time until the first arrival, we can start talking about the time until the second arrival, and so on. How do we define these? After the first arrival happens, we're going to have a sequence of time slots with no arrivals, and then the next arrival is going to happen.

So we call this time that elapses-- or number of time slots after the first arrival until the next one-- we call it T2. This is the second inter-arrival time, that is, time between arrivals. Once this arrival has happened, then we wait and see how many more it takes until the third arrival. And we call this time here, T3.

We're interested in the time of the k-th arrival, which is going to be just the sum of the first k inter-arrival times. So for example, let's say Y3 is the time that the third arrival comes. Y3 is just the sum of T1, plus T2, plus T3. So we're interested in this random variable, Y3, and it's the sum of inter-arrival times. To understand what kind of random variable it is, I guess we should understand what kind of random variables these are going to be.

So what kind of random variable is T2? Your sister is doing her coin flips until a success is observed for the first time. Based on that information about what has happened so far, you are called into the room. And you start watching until a success is observed again.

So after you start watching, what you have is just a sequence of independent Bernoulli trials. So each one of these has probability p of being a success. The time it's going to take until the first success, this number, T2, is going to be again just another geometric random variable. It's as if the process just started. After you are called into the room, you have no foresight, you don't have any information about the future, other than the fact that these are going to be independent Bernoulli trials.

So T2 itself is going to be geometric with the same parameter p. And then you can continue the arguments and argue that T3 is also geometric with the same parameter p. Furthermore, whatever happened, how long it took until you were called in, it doesn't change the statistics about what's going to happen in the future. So whatever happens in the future is independent from the past. So T1, T2, and T3 are independent random variables.

So conclusion is that the time until the third arrival is the sum of 3 independent geometric random variables, with the same parameter. And this is true more generally. The time until the k-th arrival is going to be the sum of k independent random variables. So in general, Yk is going to be T1 plus Tk, where the Ti's are geometric, with the same parameter p, and independent.

So now what's more natural than trying to find the distribution of the random variable Y_k ? How can we find it? So I fixed k for you. Let's say k is 100. I'm interested in how long it takes until 100 customers arrive. How can we find the distribution of Y_k ?

Well one way of doing it is to use this lovely convolution formula. Take a geometric, convolve it with another geometric, you get something. Take that something that you got, convolve it with a geometric once more, do this 99 times, and this gives you the distribution of Y_k . So that's definitely doable, and it's extremely tedious.

Let's try to find the distribution of Y_k using a shortcut. So the probability that Y_k is equal to t . So we're trying to find the PMF of Y_k . k has been fixed for us. And we want to calculate this probability for the various values of t , because this is going to give us the PMF of Y_k . OK.

What is this event? What does it take for the k -th arrival to be at time t ? For that to happen, we need two things. In the first $t - 1$ slots, how many arrivals should we have gotten? $k - 1$. And then in the last slot, we get one more arrival, and that's the k -th one. So this is the probability that we have $k - 1$ arrivals in the time interval from 1 up to t . And then, an arrival at time t .

That's the only way that it can happen, that the k -th arrival happens at time t . We need to have an arrival at time t . And before that time, we need to have exactly $k - 1$ arrivals. Now this is an event that refers-- $t - 1$. In the previous time slots we had exactly $k - 1$ arrivals. And then at the last time slot we get one more arrival.

Now the interesting thing is that this event here has to do with what happened from time 1 up to time $t - 1$. This event has to do with what happened at time t . Different time slots are independent of each other. So this event and that event are independent.

So this means that we can multiply their probabilities. So take the probability of this. What is that? Well probability of having a certain number of arrivals in a certain number of time slots, these are just the binomial probabilities.

So this is, out of $t - 1$ slots, to get exactly $k - 1$ arrivals, p to the $k - 1$, $(1-p)$ to the $t - 1 - (k - 1)$, this gives us $t - k$. And then we multiply with this probability, the probability of an arrival, at time t is equal to p . And so this is the formula for the PMF of the number-- of the time it takes until the k -th arrival happens.

Does it agree with the formula in your handout? Or its not there? It's not there. OK.

Yeah. OK. So that's the formula and it is true for what values of t ? [INAUDIBLE]. It takes at least k time slots in order to get k arrivals, so this formula should be true for k larger than or equal to t . For t larger than or equal to k .

All right. So this gives us the PMF of the random variable Y_k . Of course, we may also be interested in the mean and variance of Y_k . But this is a lot easier. Since Y_k is the sum of independent random variables, the expected value of Y_k is going to be just k times the expected value of your typical t .

So the expected value of Y_k is going to be just k times $1/p$, which is the mean of the geometric. And similarly for the variance, it's going to be k times the variance of a geometric. So we have everything there is to know about the distribution of how long it takes until the first arrival comes.

OK. Finally, let's do a few more things about the Bernoulli process. It's interesting to talk about several processes at the time. So in the situation here of splitting a Bernoulli process is where you have arrivals that come to a server. And that's a picture of which slots get arrivals. But actually maybe you have two servers. And whenever an arrival comes to the system, you flip a coin and with some probability, q , you send it to one server. And with probability $1-q$, you send it to another server.

So there is a single arrival stream, but two possible servers. And whenever there's an arrival, you either send it here or you send it there. And each time you decide where you send it by flipping an independent coin that has its own bias q . The coin flips that decide where do you send it are assumed to be independent from the arrival process itself.

So there's two coin flips that are happening. At each time slot, there's a coin flip that decides whether you have an arrival in this process here, and that coin flip is with parameter p . And if you have something that arrives, you flip another coin with probabilities q , and $1-q$, that decides whether you send it up there or you send it down there.

So what kind of arrival process does this server see? At any given time slot, there's probability p that there's an arrival here. And there's a further probability q that this arrival gets sent up there. So the probability that this server sees an arrival at any given time is p times q . So this process here is going to be a Bernoulli process, but with a different parameter, p times q . And this one down here, with the same argument, is going to be Bernoulli with parameter p times $(1-q)$.

So by taking a Bernoulli stream of arrivals and splitting it into two, you get two separate Bernoulli processes. This is going to be a Bernoulli process, that's going to be a Bernoulli process. Well actually, I'm running a little too fast. What does it take to verify that it's a Bernoulli process? At each time slot, it's a 0 or 1. And it's going to be a 1, you're going to see an arrival with probability p times q .

What else do we need to verify, to be able to tell-- to say that it's a Bernoulli process? We need to make sure that whatever happens in this process, in different time slots, are statistically independent from each other. Is that property true? For example, what happens in this time slot whether you got an arrival or not, is it independent from what happened at that time slot?

The answer is yes for the following reason. What happens in this time slot has to do with the coin flip associated with the original process at this time, and the coin flip that decides where to send things. What happens at that time slot has to do with the coin flip here, and the additional coin flip that decides where to send it if something came.

Now all these coin flips are independent of each other. The coin flips that determine whether we have an arrival here is independent from the coin flips that determined whether we had an arrival

there. And you can generalize this argument and conclude that, indeed, every time slot here is independent from any other time slot. And this does make it a Bernoulli process.

And the reason is that, in the original process, every time slot is independent from every other time slot. And the additional assumption that the coin flips that we're using to decide where to send things, these are also independent of each other. So we're using here the basic property that functions of independent things remain independent.

There's a converse picture of this. Instead of taking one stream and splitting it into two streams, you can do the opposite. You could start from two streams of arrivals. Let's say you have arrivals of men and you have arrivals of women, but you don't care about gender. And the only thing you record is whether, in a given time slot, you had an arrival or not.

Notice that here we may have an arrival of a man and the arrival of a woman. We just record it with a 1, by saying there was an arrival. So in the merged process, we're not keeping track of how many arrivals we had total. We just record whether there was an arrival or not an arrival.

So an arrival gets recorded here if, and only if, one or both of these streams had an arrival. So that we call a merging of two Bernoulli-- of two processes, of two arrival processes. So let's make the assumption that this arrival process is independent from that arrival process.

So what happens at the typical slot here? I'm going to see an arrival, unless none of these had an arrival. So the probability of an arrival in a typical time slot is going to be 1 minus the probability of no arrival. And the event of no arrival corresponds to the first process having no arrival, and the second process having no arrival. So there's no arrival in the merged process if, and only if, there's no arrival in the first process and no arrival in the second process.

We're assuming that the two processes are independent and that's why we can multiply probabilities here. And then you can take this formula and it simplifies to $p + q$, minus p times q .

So each time slot of the merged process has a certain probability of seeing an arrival. Is the merged process a Bernoulli process? Yes, it is after you verify the additional property that different slots are independent of each other.

Why are they independent? What happens in this slot has to do with that slot, and that slot down here. These two slots-- so what happens here, has to do with what happens here and there. What happens in this slot has to do with whatever happened here and there.

Now, whatever happens here and there is independent from whatever happens here and there. Therefore, what happens here is independent from what happens there. So the independence property is preserved. The different slots of this merged process are independent of each other. So the merged process is itself a Bernoulli process.

So please digest these two pictures of merging and splitting, because we're going to revisit them in continuous time where things are little subtler than that. OK. Good luck on the exam and see you in a week.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

LECTURE 13

The Bernoulli process

- **Readings:** Section 6.1

Lecture outline

- Definition of Bernoulli process
- Random processes
- Basic properties of Bernoulli process
- Distribution of interarrival times
- The time of the k th success
- Merging and splitting

The Bernoulli process

- A sequence of independent Bernoulli trials
- At each trial, i :
 - $P(\text{success}) = P(X_i = 1) = p$
 - $P(\text{failure}) = P(X_i = 0) = 1 - p$
- Examples:
 - Sequence of lottery wins/losses
 - Sequence of ups and downs of the Dow Jones
 - Arrivals (each second) to a bank
 - Arrivals (at each time slot) to server

Random processes

- First view:
sequence of random variables X_1, X_2, \dots
- $E[X_t] =$
- $\text{Var}(X_t) =$
- Second view:
what is the right sample space?
- $P(X_t = 1 \text{ for all } t) =$
- Random processes we will study:
 - Bernoulli process
(memoryless, discrete time)
 - Poisson process
(memoryless, continuous time)
 - Markov chains
(with memory/dependence across time)

Number of successes S in n time slots

- $P(S = k) =$
- $E[S] =$
- $\text{Var}(S) =$

Interarrival times

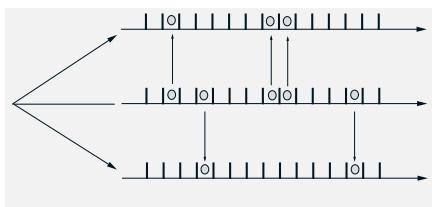
- T_1 : number of trials until first success
 - $P(T_1 = t) =$
 - Memoryless property
 - $E[T_1] =$
 - $\text{Var}(T_1) =$
- If you buy a lottery ticket every day, what is the distribution of the length of the first string of losing days?

Time of the k th arrival

- Given that first arrival was at time t i.e., $T_1 = t$: additional time, T_2 , until next arrival
 - has the same (geometric) distribution
 - independent of T_1
- Y_k : number of trials to k th success
 - $E[Y_k] =$
 - $\text{Var}(Y_k) =$
 - $P(Y_k = t) =$

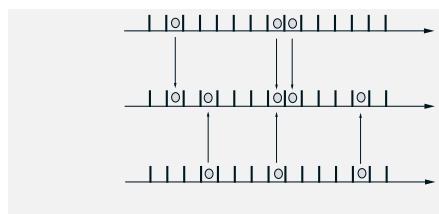
Splitting of a Bernoulli Process

(using independent coin flips)



yields Bernoulli processes

Merging of Indep. Bernoulli Processes



yields a Bernoulli process
(collisions are counted as one arrival)

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Recitation 14
October 26, 2010

1. You are visiting the rainforest, but unfortunately your insect repellent has run out. As a result, at each second, a mosquito lands on your neck with probability 0.5. If one lands, with probability 0.2 it bites you, and with probability 0.8 it never bothers you, independently of other mosquitoes.
 - (a) What is the expected time between successive mosquito bites? What is the variance of the time between successive mosquito bites?
 - (b) In addition, a tick lands on your neck with probability 0.1. If one lands, with probability 0.7 it bites you, and with probability 0.3, it never bothers you, independently of other ticks and mosquitoes. Now, what is expected time between successive bug bites? What is the variance of the time between successive bug bites?
2. Al performs an experiment comprising a series of independent trials. On each trial, he simultaneously flips a set of three fair coins.
 - (a) Given that Al has just had a trial with 3 *tails*, what is the probability that both of the next two trials will also have this result?
 - (b) Whenever all three coins land on the same side in any given trial, Al calls the trial a success.
 - i. Find the PMF for K , the number of trials up to, but *not* including, the second success.
 - ii. Find the expectation and variance of M , the number of tails that occur *before* the first success.
 - (c) Bob conducts an experiment like Al's, except that he uses 4 coins for the first trial, and then he obeys the following rule: Whenever all of the coins land on the same side in a trial, Bob permanently removes one coin from the experiment and continues with the trials. He follows this rule until the *third* time he removes a coin, at which point the experiment ceases. Find $E[N]$, where N is the number of trials in Bob's experiment.
3. Suppose there are n papers in a drawer. You draw a paper and sign it, and then, instead of filing it away, you place the paper back into the drawer. If any paper is equally likely to be drawn each time, independent of all other draws, what is the expected number of papers that you will draw before signing all n papers? You may leave your answer in the form of a summation.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Recitation 14 Solutions
October 26, 2010

1. (a) Let $X = (\text{time between successive mosquito bites}) = (\text{time until the next mosquito bite})$.

The mosquito bites occur according to a Bernoulli process with parameter $p = 0.5 \cdot 0.2 = 0.1$. X is a geometric random variable, so, $\mathbf{E}[X] = \frac{1}{p} = \frac{1}{0.1} = 10$.

$$\text{var}(X) = \frac{1-p}{p^2} = \frac{1-0.1}{0.1^2} = 90.$$

- (b) Mosquito bites occur according to a Bernoulli process with parameter $p = 0.1$. Tick bites occur according to another independent Bernoulli process with parameter $q = 0.1 \cdot 0.7 = 0.07$. Bug bites (mosquito or tick) occur according to a merged Bernoulli process from the mosquito and tick processes. Therefore, the probability of success at any time point for the merged Bernoulli process is $r = p + q - pq = 0.1 + 0.07 - 0.1 \cdot 0.07 = 0.163$. Let Y be the time between successive bug bites. As before, Y is a geometric random variable, so $\mathbf{E}[Y] = \frac{1}{r} = \frac{1}{0.163} \approx 6.135$.

$$\text{var}(Y) = \frac{1-r}{r^2} = \frac{1-0.163}{0.163^2} \approx 31.503$$

2. (a) In this case, since the trials are independent, the given information is irrelevant.

$$\mathbf{P}(\text{next 2 trials result in 3 tails}) = \left(\frac{1}{8}\right)^2 = \frac{1}{64}.$$

- (b) i. The second order Pascal PMF for random variable N , as defined in the text, is the probability of the second success comes on the n^{th} trial. Thus, the random variable, K , is a shifted version of the second order Pascal PMF, i.e. $K = N - 1$. So, the probability that 1 success comes in the first k trials, where the next trial will result in the second success, can be expressed as:

$$p_K(k) = \binom{k}{1} \left(\frac{1}{4}\right)^2 \left(\frac{3}{4}\right)^{k-1}, \quad k \geq 1.$$

- ii. The number of tails before the first success, M , can be written as a random sum:

$$M = X_1 + X_2 + \cdots + X_N,$$

where X_i is the number of tails that occur on (unsuccessful) trial i , and N is the number of unsuccessful trials (i.e. trials before the first success). We notice that X is equally likely to be either 1 or 2, and that N is a shifted geometric: $N = R - 1$, where R is a geometric random variable with parameter $\frac{1}{4}$. Now we can apply our random sum formulae.

$$\mathbf{E}[M] = \mathbf{E}[X]\mathbf{E}[N] = \left(\frac{3}{2}\right)(4-1) = \frac{9}{2}$$

$$\text{var}(M) = \mathbf{E}[N]\text{var}(X) + (\mathbf{E}[X])^2\text{var}(N) = (4-1)\left(\frac{1}{4}\right) + \left(\frac{3}{2}\right)^2(12) = \frac{111}{4}.$$

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
 (Fall 2010)

- (c) N , the number of trials in Bob's experiment, can be expressed as the sum of 3 independent random variables, X , Y , and Z . X is the number of trials until Bob removes the first coin, Y the number of additional trials until he removes the second coin, and Z the additional number until he removes the third coin. We see that X is a geometric random variable with parameter $\frac{1}{8}$, Y is geometric with parameter $\frac{1}{4}$, and Z geometric with parameter $\frac{1}{2}$. Hence,

$$E[N] = E[X] + E[Y] + E[Z] = 8 + 4 + 2 = 14.$$

3. Let M be the total number of draws you make until you have signed all n papers. Let T_i be the number of draws you make until drawing the next unsigned paper after having signed i papers. Then $M = T_0 + \dots + T_{n-1}$.

We can view the process of selecting the next unsigned paper after having signed i papers as a sequence of independent Bernoulli trials with probability of success $p_i = \frac{n-i}{n}$, since there are $n-i$ unsigned papers out of a total of n papers and receiving any paper is equally likely in a particular draw. The PMF governing the number of attempts we make until we succeed in drawing the next unsigned paper after having signed i papers is geometric. More concretely, the probability that it takes k tries to draw the next unsigned paper after having signed i papers is

$$\mathbf{P}(T_i = k) = (1 - p_i)^{k-1} p_i.$$

With this model, the expected value of M , the number of draws you make until you sign all n papers is:

$$\mathbf{E}[M] = \mathbf{E} \left[\sum_{i=0}^{n-1} T_i \right] = \sum_{i=0}^{n-1} \mathbf{E}[T_i] = \sum_{i=0}^{n-1} \frac{n}{n-i} = n \sum_{k=1}^n \frac{1}{k}.$$

For large n , this is on the order of: $n \int_1^n \frac{1}{x} dx = n \log n$.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Tutorial 7
October 28/29, 2010

1. Alice and Bob alternate playing at the casino table. (Alice starts and plays at odd times $i = 1, 3, \dots$; Bob plays at even times $i = 2, 4, \dots$) At each time i , the net gain of whoever is playing is a random variable G_i with the following PMF:

$$p_G(g) = \begin{cases} \frac{1}{3} & g = -2, \\ \frac{1}{3} & g = 1, \\ \frac{1}{6} & g = 3, \\ 0 & \text{otherwise} \end{cases}$$

Assume that the net gains at different times are independent. We refer to an outcome of -2 as a “loss.”

- (a) They keep gambling until the first time where a loss by Bob immediately follows a loss by Alice. Write down the PMF of the total number of rounds played. (A round consists of two plays, one by Alice and then one by Bob.)
- (b) Write down the PMF for Z , defined as the time at which Bob has his third loss.
- (c) Let N be the number of rounds until each one of them has won at least once. Find $\mathbf{E}[N]$.

2. Problem 6.6, page 328 in text.

Sum of a geometric number of independent geometric random variables

Let $Y = X_1 + \dots + X_N$, where the random variable X_i are geometric with parameter p , and N is geometric with parameter q . Assume that the random variables N, X_1, X_2, \dots are independent. Show that Y is geometric with parameter pq . *Hint:* Interpret the various random variables in terms of a split Bernoulli process.

3. A train bridge is constructed across a wide river. Trains arrive at the bridge according to a Poisson process of rate $\lambda = 3$ per day.

- If a train arrives on day 0, find the probability that there will be no trains on days 1, 2, and 3.
- Find the probability that the next train to arrive after the first train on day 0, takes more than 3 days to arrive.
- Find the probability that no trains arrive in the first 2 days, but 4 trains arrive on the 4th day.
- Find the probability that it takes more than 2 days for the 5th train to arrive at the bridge.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Tutorial 7: Solutions

1. (a) For each round, the probability that both Alice and Bob have a loss is $\frac{1}{3} \cdot \frac{1}{3} = \frac{1}{9}$. Let random variable X represent the total number of rounds played until the first time where they both have a loss. Then X is a geometric random variable with parameter $p = 1/9$ and has the following PMF.

$$p_X(x) = (1 - p)^{x-1} p = \left(\frac{8}{9}\right)^{x-1} \left(\frac{1}{9}\right), \quad x = 1, 2, \dots$$

- (b) First, consider the number of games, K_3 Bob played until his third loss. Random variable K_3 is a Pascal random variable and has the following PMF.

$$p_{K_3}(k) = \binom{k-1}{3-1} \left(\frac{1}{3}\right)^3 \left(\frac{2}{3}\right)^{k-3} \quad k = 3, 4, 5, \dots$$

In this question, we are interested in another random variable Z defined as the time at which Bob has his third loss. Note that $Z = 2K_3$. By changing variables, we obtain

$$p_Z(z) = \binom{\frac{z}{2}-1}{3-1} \left(\frac{1}{3}\right)^3 \left(\frac{2}{3}\right)^{\frac{z}{2}-3} \quad z = 6, 8, 10, \dots$$

- (c) Let A be the event that Alice wins, and Let B be the event that Bob wins. The event $A \cup B$ is then the event that either A wins or B wins or both A and B win, and the event $A \cap B$ is the event that both A and B win. Suppose we observe this gambling process, and let U be a random variable indicating the number of rounds we see until at least one of them wins. Random variable U is a geometric random variable with parameter $p = P(A \cup B) = 1 - \frac{1}{3} \cdot \frac{1}{3}$.

Consider another random variable V representing the number of additional rounds we have to observe until the other wins. If both Alice and Bob win at the U th round, then $V = 0$. This occurs with probability $P(A \cap B | A \cup B) = \frac{2}{9}$. If Alice wins the U th round, then the time V until Bob wins is a geometric random variable with parameter $p = 1/2 + 1/6 = 2/3$. This occurs with probability $P(A \cap B^c | A \cup B) = \frac{1}{3} \cdot \frac{2}{3}$. Likewise, if Bob wins the U th round, then the time V until Alice wins is a geometric random variable with parameter $p = 1/2 + 1/6 = 2/3$. This occurs with probability $P(B \cap A^c | A \cup B) = \frac{1}{3} \cdot \frac{2}{3}$. The number of rounds until each one of them has won at least once, N is

$$N = U + V$$

The expectation of N is then:

$$\begin{aligned} E[N] &= E[U] + E[V] \\ &= \frac{1}{\frac{8}{9}} + 0 \cdot P(A \cap B | A \cup B) + \frac{1}{\frac{2}{3}} P(A | A \cup B) + \frac{1}{\frac{2}{3}} P(B | A \cup B) \\ &= 9/8 + \frac{3}{2} \frac{\frac{1}{2} \frac{2}{3}}{\frac{8}{9}} + \frac{3}{2} \frac{\frac{1}{2} \frac{2}{3}}{\frac{8}{9}} \\ &= 15/8 \end{aligned}$$

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
 (Fall 2010)

There is another approach to this problem. Consider the following partition.

- A_1 : both win first round
- A_2 : Only Alice wins first round
- A_3 : Only Bob wins first round
- A_4 : both lose first round

Event A_1 occurs with probability $\frac{2}{3} \cdot \frac{2}{3}$. Event A_2 occurs with probability $\frac{2}{3} \cdot \frac{1}{3}$. Event A_3 occurs with probability $\frac{1}{3} \cdot \frac{2}{3}$. Event A_4 occurs with probability $\frac{1}{3} \cdot \frac{1}{3}$. When event A_2 (A_3) occurs, the distribution on the time until Bob (Alice) wins is a geometric random variable with mean $\frac{1}{\frac{2}{3}} = \frac{3}{2}$. When event A_4 occurs, the additional time until Alice and Bob win is distributed identically to that at time 0 by the fresh-start property. By the total expectation theorem,

$$\begin{aligned} E[N] &= E[N|A_1]P(A_1) + E[N|A_2]P(A_2) + E[N|A_3]P(A_3) + E[N|A_4]P(A_4) \\ &= 1 \cdot \left(\frac{2}{3} \cdot \frac{2}{3}\right) + \left(1 + \frac{1}{\frac{2}{3}}\right) \cdot \left(\frac{1}{3} \cdot \frac{2}{3}\right) + \left(\frac{1}{3} \cdot \frac{2}{3}\right) + (1 + E[N]) \cdot \left(\frac{1}{3} \cdot \frac{1}{3}\right) \end{aligned}$$

Solving for $E[N]$, we get $E[N] = \frac{15}{8}$.

2. Problem 6.6, page 328 in text. See text for solutions.
3. (a) The number of trains arriving on days 1, 2, and 3 is independent of the number of trains arriving on day 0. Let N denote the total number of trains that arrive on days 1, 2, and 3. Then N is a Poisson random variable with parameter $3\lambda = 9$, and we have

$$\begin{aligned} P(\text{no train on days 1,2,3} \mid \text{one train on day 1}) &= P(\text{no train on days 1,2,3}) \\ &= P(N = 0) \\ &= \frac{e^{-9}9^0}{0!} \\ &= e^{-9}. \end{aligned}$$

- (b) The event that the next arrival is more than three days after the train arrival on day 0 is the same as the event that there are zero arrivals in the three days after the train arrival on day 0. Therefore the required probability is the same as that found in part (a), namely, e^{-9} .
- (c) The number of trains arriving in the first 2 days is independent of the number of trains arriving on day 4. Therefore, we have

$$\begin{aligned} P(\text{0 trains in first 2 days and 4 trains on day 4}) &= P(\text{0 trains in first 2 days}) \cdot P(\text{4 trains on day 4}) \\ &= e^{-2\lambda} \frac{(2\lambda)^0}{0!} \cdot e^{-\lambda} \frac{\lambda^4}{4!} \\ &= e^{-9} \frac{3^4}{4!}. \end{aligned}$$

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Fall 2010)

- (d) The event that it takes more than 2 days for the 5th arrival is equivalent to the event that there are at most 4 arrivals in the first 2 days. Therefore the required probability is equal to

$$\begin{aligned}\sum_{k=0}^4 P(\text{exactly } k \text{ arrivals in first 2 days}) &= \sum_{k=0}^4 e^{-2\lambda} \frac{(2\lambda)^k}{k!} \\ &= e^{-2\lambda} \left(\frac{(2\lambda)^0}{0!} + \frac{(2\lambda)^1}{1!} + \frac{(2\lambda)^2}{2!} + \frac{(2\lambda)^3}{3!} + \frac{(2\lambda)^4}{4!} \right) \\ &= e^{-6}(1 + 6 + 18 + 36 + 54) \\ &= 115e^{-6}.\end{aligned}$$

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013 Transcript – Recitation: Competing Exponentials

Hi, in this problem, we're going to look at competing exponential. So we have three exponential random variables, X with parameter lambda, Y with parameters mu, and Z with parameter nu. And we want to calculate some probability. And the probability that we want to calculate is the probability that X is less than Y is less than Z.

Now we can reinterpret this as 3 plus Poisson processes. Because the link between exponentials and Poisson processes is that the inter-arrival time of Poisson processes are exponentially distributed. So you can think of X being the time until the first arrival in a Poisson process with parameter lambda. And same thing for Y is the first arrival time of a Poisson process with parameter mu. The same thing for Z and nu.

And so in that interpretation, X less than Y less than Z, you could interpret as a race, meaning that X finishes first followed by Y and then Z comes in last. So with that interpretation, let's see if we can calculate what this probability is. We can rewrite this probability as a combination of two things occurring. One is that X is less than the minimum of Y and Z. And then the other is that Y is less than Z.

So what is this first event mean? This first event means that X comes in first. And it doesn't matter whether Y comes in second or Z comes in second. So we first say that X has to come in first which means it has to beat the better of Y and Z. And then that combined with the fact that Y does better than Z is the same thing as saying that X is first, Y is second, and Z is third.

And now, let's try to argue using Poisson processes, that these two events are actually independent. So this event occurring means that X is smaller than Y and Z. So let's take these Poisson processes, and because these random variables are assumed to be independent, these are independent Poisson processes. So we can merge them. So let's merge these two. And we'll get a Poisson process that has rate mu plus nu.

And we can also merge this first one and that one. And we'll get another Poisson process with predator lambda plus mu plus nu. So in that context, what does it mean that X is less than the minimum Y and Z? It just means that in this merged process, the first arrival came from the X process. In that case, if that's true, then X is less than minimum Y and Z. Well, let's say that event does occur that the first arrival is from the X process. Now we're interested in what the order of the second two arrivals are. Is it Y first and then Z? Or Z first and then Y?

Well, it doesn't matter because of the fresh start property. Because after this arrival comes, and say it is from the X process, the Poisson processes start anew, and they're still independent. And so what happens after that is independent of what happened here, when X arrived. And so whether Y came first followed by Z, or Z came first followed by Y is independent of what happened here. And so because of that, these two events are independent, and so when we have

the probability of the intersection of two independent events, we can write that as the product of those two probabilities.

Now, what is the probability of this first event? The probability that X is less than the minimum Y and Z ? Well, we just said that that corresponds to the first arrival of this merge process being from the X process. Well, that probability is λ over $\lambda + \mu + \nu$. So it's equal to this ratio where the process that you're interested in, its rate comes in the numerator. And then the merged rate is on the denominator,

And what about the second one? What's the probability that Y is less than Z ? Well, let's go now to this merge process where we merged just the Y and Z processes and see which one comes first. Well, in that case what we want to know is in this merge process, what is the probability that the first arrival came from the Y process? Well, analogously, that probability is going to be μ over $\mu + \nu$. And that gives us our answer.

And so we see that what looked like a pretty complex calculation when we reinterpreted it in terms of Poisson processes, it becomes relatively easy to solve. But this still seems like a complicated expression. So let's try to check to see whether it actually makes sense. So one way to do that is to look at a specific example of the choice of λ , μ , and ν , and see if it actually makes sense.

So one example is suppose that all three of these parameters are the same. Well, if they're all the same then this probability, the first part becomes $1/3$. And the second one is $1/2$. And so if all three parameters are the same, probability becomes $1/6$. And let's see if that makes sense. If all three parameters are the same, that means that these rates, these arrival rates are all the same. And what that means is that any three ordering of these three arrivals is as likely as any other ordering. And what we're interested in is the probability of one's particular ordering happening which is X first then Y then Z .

But if everything is symmetric then any of the orderings is as likely as any other one. And how many orderings are there? Well, there's three choices for who comes in first. Two for who comes in second. And one for who comes in last. So there's a total of six possible orders in which these three contestants, if you think of it that way, could finish this race. And out of none of those, we want the probability that one of those outcomes happens. And so the probability should be $1/6$. And that's what our formula tells us. So as I said, in this problem, we saw how to reinterpret exponentials in the context of Poisson processes that helped us solve a--

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013

Transcript – Lecture 14

The following content is provided under a Creative Commons license. Your support will help MIT OpenCourseWare continue to offer high quality educational resources for free. To make a donation or view additional materials from hundreds of MIT courses, visit MIT OpenCourseWare at ocw.mit.edu.

PROFESSOR: So last time we started talking about random processes. A random process is a random experiment that evolves over time. And conceptually, it's important to realize that it's a single probabilistic experiment that has many stages. Actually, it has an infinite number of stages. And we discussed the simplest random process there is, the Bernoulli process, which is nothing but the sequence of Bernoulli trials-- an infinite sequence of Bernoulli trials. For example, flipping a coin over and over.

Once we understand what's going on with that process, then what we want is to move into a continuous time version of the Bernoulli process. And this is what we will call the Poisson process. And for the Poisson process, we're going to do exactly the same things that we did for the Bernoulli process. That is, talk about the number of arrivals during a given time period, and talk also about the time between consecutive arrivals, and for the distribution of inter-arrival times.

So let's start with a quick review of what we discussed last time. First, a note about language. If you think of coin tosses, we then talk about heads and tails. If you think of these as a sequence of trials, you can talk about successes and failures. The language that we will be using will be more the language of arrivals. That is, if in a given slot you have a success, you say that something arrived. If you have a failure, nothing arrived. And that language is a little more convenient and more natural, especially when we talk about continuous time-- to talk about arrivals instead of successes.

But in any case, for the Bernoulli process let's keep, for a little bit, the language of successes. Whereas working in discrete time, we have time slots. During each time slot, we have an independent Bernoulli trial. There is probability p of having a success. Different slots are independent of each other. And this probability p is the same for any given time slot.

So for this process we will discuss the one random variable of interest, which is the following. If we have n time slots, or n trials, how many arrivals will there be? Or how many successes will there be? Well, this is just given by the binomial PMF. Number of successes in n trials is a random variable that has a binomial PMF, and we know what this is.

Then we talked about inter-arrival times. The time until the first arrival happens has a geometric distribution. And we have seen that from some time ago. Now if you start thinking about the time until k arrivals happen, and we denote that by Y_k , this is the time until the first arrival

happens. And then after the first arrival happens, you have to wait some time until the second arrival happens, and so on. And then the time from the $(k - 1)$ th arrival, until arrival number k .

The important thing to realize here is that because the process has a memorylessness property, once the first arrival comes, it's as if we're starting from scratch and we will be flipping our coins until the next arrival comes. So the time it will take until the next arrival comes will also be a geometric random variable. And because different slots are independent, whatever happens after the first arrival is independent from whatever happened before. So T_1 and T_2 will be independent random variables. And similarly, all the way up to T_k .

So the time until the k -th arrival is a sum of independent geometric random variables, with the same parameter p . And we saw last time that we can find the probability distribution of Y_k . The probability that Y_k takes a value of t is equal to-- there's this combinatorial factor here, and then you get p to the k , $(1-p)$ to the $(t-k)$, and this formula is true for t equal to k , $k+1$, and so on. And this distribution has a name. It's called the Pascal PMF.

So this is all there is to know about the Bernoulli process. One important comment is to realize what exactly this memorylessness property is saying. So I discussed it a little bit last time. Let me reiterate it. So we have a Bernoulli process, which is a sequence of Bernoulli trials. And these are $(0,1)$ random variables that keep going on forever.

So someone is watching this movie of Bernoulli trials B_t . And at some point, they say they think, or something interesting has happened, why don't you come in and start watching? So at some time t , they tell you to come in and start watching. So what you will see once you come in will be this future trials.

So actually what you will see is a random process, whose first random variable is going to be the first one that you see, $B_{(t+1)}$. The second one is going to be this, and so on. So this is the process that's seen by the person who's asked to come in and start watching at that time. And the claim is that this process is itself a Bernoulli process, provided that the person who calls you into the room does not look into the future. The person who calls you into the room decides to call you in only on the basis of what they have seen so far.

So for example, who calls you into the room might have a rule that says, as soon as I see a sequence of 3 heads, I ask the other person to come in. So if they use that particular rule, it means that when you're called in, the previous 3 were heads. But this doesn't give you any information about the future. And so the future ones will be just independent Bernoulli trials.

If on the other hand, the person who calls you in has seen the movie before and they use a rule, such as, for example, I call you in just before 3 heads show up for the first time. So the person calls you in based on knowledge that these two would be three heads. If they have such foresight-- if they can look into the future-- then X_1 , X_2 , X_3 , they're certain to be three heads, so they do not correspond to random independent Bernoulli trials.

So to rephrase this, the process is memoryless. It does not matter what has happened in the past. And that's true even if you are called into the room and start watching at a random time, as long as that random time is determined in a causal way on the basis of what has happened so far.

So you are called into the room in a causal manner, just based on what's happened so far. What you're going to see starting from that time will still be a sequence of independent Bernoulli trials. And this is the argument that we used here, essentially, to argue that this T_2 is an independent random variable from T_1 .

So a person is watching the movie, sees the first success. And on the basis of what they have seen-- they have just seen the first success-- they ask you to come in. You come in. What you're going to see is a sequence of Bernoulli trials. And you wait this long until the next success comes in. What you see is a Bernoulli process, as if the process was just starting right now. And that convinces us that this should be a geometric random variable of the same kind as this one, as independent from what happened before.

All right. So this is pretty much all there is to know about the Bernoulli process. Plus the two things that we did at the end of the last lecture where we merge two independent Bernoulli processes, we get a Bernoulli process. If we have a Bernoulli process and we split it by flipping a coin and sending things one way or the other, then we get two separate Bernoulli processes. And we see that all of these carry over to the continuous time. And our task for today is basically to work these continuous time variations.

So the Poisson process is a continuous time version of the Bernoulli process. Here's the motivation for considering it a Bernoulli process. So you have that person whose job is to sit outside the door of a bank. And they have this long sheet, and for every one second slot, they mark an X if a person came in, or they mark something else if no one came in during that slot.

Now the bank manager is a really scientifically trained person and wants very accurate results. So they tell you, don't use one second slots, use milliseconds slots. So you have all those slots and you keep filling if someone arrived or not during that slot. Well then you come up with an idea. Why use millisecond slots and keep putting crosses or zero's into each slot? It's much simpler if I just record the exact times when people came in. So time is continuous.

I don't keep doing something at every time slot. But instead of the time axis, I mark the times at which customers arrive. So there's no real need for slots. The only information that you want is when did we have arrivals of people. And we want to now model a process of this kind happening in continuous time, that has the same flavor, however, as the Bernoulli process. So that's the model we want to develop.

OK. So what are the properties that we're going to have? First, we're going to assume that intervals over the same length behave probabilistically in an identical fashion. So what does that mean? Think of an interval of some given length. During the interval of that length, there's going to be a random number of arrivals. And that random number of arrivals is going to have a probability distribution. So that probability distribution-- let's denote it by this notation.

We fix t , we fix the duration. So this is fixed. And we look at the different k 's. The probability of having 0 arrivals, the probability of 1 arrival, the probability of 2 arrivals, and so on. So this thing is essentially a PMF. So it should have the property that the sum over all k 's of this $P_{(k, \tau)}$ should be equal to 1.

Now, hidden inside this notation is an assumption of time homogeneity. That is, this probability distribution for the number of arrivals only depends on the length of the interval, but not the exact location of the interval on the time axis.

That is, if I take an interval of length τ , and I ask about the number of arrivals in this interval. And I take another interval of length τ , and I ask about the number of arrivals during that interval. Number of arrivals here, and number of arrivals there have the same probability distribution, which is denoted this way.

So the statistical behavior of arrivals here is the same as the statistical behavioral of arrivals there. What's the relation with the Bernoulli process? It's very much like the assumption-- the Bernoulli process-- that in different slots, we have the same probability of success. Every slot looks probabilistically as any other slot.

So similarly here, any interval of length τ looks probabilistically as any other interval of length τ . And the number of arrivals during that interval is a random variable described by these probabilities. Number of arrivals here is a random variable described by these same probabilities. So that's our first assumption.

Then what else? In the Bernoulli process we had the assumption that different time slots were independent of each other. Here we do not have time slots, but we can still think in a similar way and impose the following assumption, that these joint time intervals are statistically independent. What does that mean?

Does a random number of arrivals during this interval, and the random number of arrivals during this interval, and the random number of arrivals during this interval-- so these are three different random variables-- these three random variables are independent of each other. How many arrivals we got here is independent from how many arrivals we got there.

So this is similar to saying that different time slots were independent. That's what we did in discrete time. The continuous time analog is this independence assumption. So for example, in particular, number of arrivals here is independent from the number of arrivals there. So these are two basic assumptions about the process.

Now in order to write down a formula, eventually, about this probability distribution-- which is our next objective, we would like to say something specific about this distribution of number of arrivals-- we need to add a little more structure into the problem.

And we're going to make the following assumption. If we look at the time interval of length δ -- and δ now is supposed to be a small number, so a picture like this-- during a very small time interval, there is a probability that we get exactly one arrival, which is λ times

delta. Delta is the length of the interval and lambda is a proportionality factor, which is sort of the intensity of the arrival process.

Bigger lambda means that a little interval is more likely to get an arrival. So there's a probability lambda times delta of 1 arrival. The remaining probability goes to 0 arrivals. And when delta is small, the probability of 2 arrivals can be approximated by 0. So this is a description of what happens during a small, tiny slot.

Now this is something that's supposed to be true in some limiting sense, when delta is very small. So the exact version of this statement would be that this is an equality, plus order of delta squared terms. So this is an approximate equality. And what approximation means is that in the limit of small deltas, the dominant terms-- the constant and the first order term are given by this.

Now when delta is very small, second order terms in delta do not matter. They are small compared to first order terms. So we ignore this. So you can either think in terms of an exact relation, which is the probabilities are given by this, plus delta squared terms. Or if you want to be a little more loose, you just write here, as an approximate equality. And the understanding is that this equality holds-- approximately becomes more and more correct as delta goes to 0.

So another version of that statement would be that if you take the limit as delta goes to 0, of p, the probability of having 1 arrival in an interval of length delta, divided by delta, this is equal to lambda. So that would be one version of an exact statement of what we are assuming here.

So this lambda, we call it the arrival rate, or the intensity of the process. And clearly, if you double lambda, then a little interval is likely -- you expect to get -- the probability of obtaining an arrival during that interval has doubled. So in some sense we have twice as intense arrival process.

If you look at the number of arrivals during delta interval, what is the expected value of that random variable? Well with probability lambda delta we get 1 arrival. And with the remaining probability, we get 0 arrivals. So it's just lambda times delta. So expected number of arrivals during a little interval is lambda times delta. So expected number of arrivals is proportional to lambda, and that's again why we call lambda the arrival rate.

If you send delta to the denominator in this equality, it tells you that lambda is the expected number of arrivals per unit time. So the arrival rate is expected number of arrivals per unit time. And again, that justifies why we call lambda the intensity of this process.

All right. So where are we now? For the Bernoulli process, the number of arrivals during a given interval of length n had the PMF that we knew it was the binomial PMF. What is the formula for the corresponding PMF for the continuous time process? Somehow we would like to use our assumptions and come up with the formula for this quantity.

So this tells us about the distribution of number of arrivals during an interval of some general length. We have made assumptions about the number of arrivals during an interval of small

length. An interval of big length is composed of many intervals of small length, so maybe this is the way to go. Take a big interval, and split it into many intervals of small length.

So we have here our time axis. And we have an interval of length tau. And I'm going to split it into lots of little intervals of length delta. So how many intervals are we going to have? The number of intervals is going to be the total time, divided by delta.

Now what happens during each one of these little intervals? As long as the intervals are small, what you have is that during an interval, you're going to have either 0 or 1 arrival. The probability of more than 1 arrival during a little interval is negligible.

So with this picture, you have essentially a Bernoulli process that consists of so many trials. And during each one of those trials, we have a probability of success, which is lambda times delta.

Different little intervals here are independent of each other. That's one of our assumptions, that these joint time intervals are independent. So approximately, what we have is a Bernoulli process. We have independence. We have the number of slots of interest. And during each one of the slots we have a certain probability of success.

So if we think of this as another good approximation of the Poisson process-- with the approximation becoming more and more accurate as delta goes to 0 -- what we should do would be to take the formula for the PMF of number of arrivals in a Bernoulli process, and then take the limit as delta goes to 0.

So in the Bernoulli process, the probability of k arrivals is n choose k, and then you have p to the k. Now in our case, we have here lambda times delta, delta is tau over n. Delta is tau over n, so p is lambda times tau divided by n. So here's our p -- Lambda tau over n -- to the power k, and then times one minus this-- this is our one minus p-- to the power n-k.

So this is the exact formula for the Bernoulli process. For the Poisson process, what we do is we take that formula and we let delta go to 0. As delta goes to 0, n goes to infinity. So that's the limit that we're taking.

On the other hand, this expression lambda times tau-- lambda times tau, what is it going to be? Lambda times tau is equal to n times p. n times p, is that what I want? No, let's see. Lambda tau is np. Yeah. So lambda tau is np.

All right. So we have this relation, lambda tau equals np. These two numbers being equal kind of makes sense. np is the expected number of successes you're going to get in the Bernoulli process. Lambda tau-- since lambda is the arrival rate and you have a total time of tau, lambda tau you can think of it as the number of expected arrivals in the Bernoulli process.

We're doing a Bernoulli approximation to the Poisson process. We take the formula for the Bernoulli, and now take the limit as n goes to infinity. Now lambda tau over n is equal to p, so it's clear what this term is going to give us. This is just p to the power k.

It will actually take a little more work than that. Now I'm not going to do the algebra, but I'm just telling you that one can take the limit in this formula here, as n goes to infinity. And that will give you another formula, the final formula for the Poisson PMF.

One thing to notice is that here you have something like $1 - \text{constant}/n$, to the power n . And you may recall from calculus a formula of this kind, that this converges to e to the minus c . If you remember that formula from calculus, then you will expect that here, in the limit, you are going to get something like $e^{-\lambda t}$. So indeed, we will get such a term.

There is some work that needs to be done to find the limit of this expression, times that expression. The algebra is not hard, it's in the text. Let's not spend more time doing this. But let me just give you the formula of what comes at the end. And the formula that comes at the end is of this form.

So what matters here is not so much the specific algebra that you will do to go from this formula to that one. It's kind of straightforward. What's important is the idea that the Poisson process, by definition, can be approximated by a Bernoulli process in which we have a very large number of slots-- n goes to infinity. Whereas we have a very small probability of success during each time slot. So a large number of slots, but tiny probability of success during each slot. And we take the limit as the slots become smaller and smaller.

So with this approximation we end up with this particular formula. And this is the so-called Poisson PMF. Now this function P here -- has two arguments. The important thing to realize is that when you think of this as a PMF, you fix t to τ . And for a fixed τ , now this is a PMF. As I said before, the sum over k has to be equal to 1. So for a given τ , these probabilities add up to 1. The formula is moderately messy, but not too messy. One can work with it without too much pain.

And what's the mean and variance of this PMF? Well what's the expected number of arrivals? If you think of this Bernoulli analogy, we know that the expected number of arrivals in the Bernoulli process is n times p . In the approximation that we're using in these procedure, n times p is the same as $\lambda \tau$. And that's why we get $\lambda \tau$ to be the expected number of arrivals. Here I'm using t instead of τ . The expected number of arrivals is λt .

So if you double the time, you expect to get twice as many arrivals. If you double the arrival rate, you expect to get twice as many arrivals. How about the formula for the variance? The variance of the Bernoulli process is np , times one minus p .

What does this go to in the limit? In the limit that we're taking, as Δ goes to zero, then p also goes to zero. The probability of success in any given slot goes to zero. So this term becomes insignificant. So this becomes n times p , which is again λt , or $\lambda \tau$.

So the variance, instead of having this more complicated formula of the variance is the Bernoulli process, here it gets simplified and it's λt . So interestingly, the variance in the Poisson process is exactly the same as the expected value. So you can look at this as just some interesting coincidence.

So now we're going to take this formula and see how to use it. First we're going to do a completely trivial, straightforward example. So 15 years ago when that example was made, email was coming at a rate of five messages per hour. I wish that was the case today. And now emails that are coming in, let's say during the day-- the arrival rates of emails are probably different in different times of the day. But if you fix a time slot, let's say 1:00 to 2:00 in the afternoon, there's probably a constant rate. And email arrivals are reasonably well modeled by a Poisson process.

Speaking of modeling, it's not just email arrivals. Whenever arrivals happen in a completely random way, without any additional structure, the Poisson process is a good model of these arrivals. So the times at which car accidents will happen, that's a Poisson processes. If you have a very, very weak light source that's shooting out photons, just one at a time, the times at which these photons will go out is well modeled again by a Poisson process. So it's completely random.

Or if you have a radioactive material where one atom at a time changes at random times. So it's a very slow radioactive decay. The time at which these alpha particles, or whatever we get emitted, again is going to be described by a Poisson process. So if you have arrivals, or emissions, that happen at completely random times, and once in a while you get an arrival or an event, then the Poisson process is a very good model for these events.

So back to emails. Get them at a rate of five messages per day, per hour. In 30 minutes this is half an hour. So what we have is that λt , total number of arrivals is-- the expected number of arrivals is-- λ is five, t is one-half, if we talk about hours. So λt is two to the 0.5.

The probability of no new messages is the probability of zero, in time interval of length t , which, in our case, is one-half. And then we look back into the formula from the previous slide, and the probability of zero arrivals is λt to the power zero, divided by zero factorial, and then an e to the λt . And you plug in the numbers that we have. λt to the zero power is one. Zero factorial is one. So we're left with e to the minus 2.5. And that number is 0.08.

Similarly, you can ask for the probability that you get exactly one message in half an hour. And that would be-- the probability of one message in one-half an hour-- is going to be λt to the first power, divided by 1 factorial, e to the minus λt , which-- as we now get the extra λt factor-- is going to be 2.5, e to the minus 2.5. And the numerical answer is 0.20. So this is how you use the PMF formula for the Poisson distribution that we had in the previous slide.

All right. So this was all about the distribution of the number of arrivals. What else did we do last time? Last time we also talked about the time it takes until the k -th arrival. OK. So let's try to figure out something about this particular distribution. We can derive the distribution of the time of the k -th arrival by using the exact same argument as we did last time.

So now the time of the k -th arrival is a continuous random variable. So it has a PDF. Since we are in continuous time, arrivals can happen at any time. So Y_k is a continuous random variable. But now let's think of a time interval of length little delta. And use our usual interpretation of PDFs. The PDF of a random variable evaluated at a certain time times delta, this is the probability that the Y_k falls in this little interval.

So as I've said before, this is the best way of thinking about PDFs. PDFs give you probabilities of little intervals. So now let's try to calculate this probability. For the k-th arrival to happen inside this little interval, we need two things. We need an arrival to happen in this interval, and we need k minus one arrivals to happen during that interval.

OK. You'll tell me, but it's possible that we might have the k minus one arrival happen here, and the k-th arrival to happen here. In principle, that's possible. But in the limit, when we take delta very small, the probability of having two arrivals in the same little slot is negligible. So assuming that no two arrivals can happen in the same mini slot, then for the k-th one to happen here, we must have k minus one during this interval.

Now because we have assumed that these joint intervals are independent of each other, this breaks down into the probability that we have exactly k minus one arrivals, during the interval from zero to t, times the probability of exactly one arrival during that little interval, which is lambda delta.

We do have a formula for this from the previous slide, which is lambda t, to the k minus 1, over k minus one factorial, times e to minus lambda t. And then lambda times delta. Did I miss something?

Yeah, OK. All right. And now you cancel this delta with that delta. And that gives us a formula for the PDF of the time until the k-th arrival. This PDF, of course, depends on the number k. The first arrival is going to happen somewhere in this range of time. So this is the PDF that it has.

The second arrival, of course, is going to happen later. And the PDF is this. So it's more likely to happen around these times. The third arrival has this PDF, so it's more likely to happen around those times.

And if you were to take k equal to 100, you might get a PDF-- it's extremely unlikely that the k-th arrival happens in the beginning, and it might happen somewhere down there, far into the future. So depending on which particular arrival we're talking about, it has a different probability distribution. The time of the 100th arrival, of course, is expected to be a lot larger than the time of the first arrival.

Incidentally, the time of the first arrival has a PDF whose form is quite simple. If you let k equal to one here, this term disappears. That term becomes a one. You're left with just lambda, e to the minus lambda. And you recognize it, it's the exponential distribution. So the time until the first arrival in a Poisson process is an exponential distribution.

What was the time of the first arrival in the Bernoulli process? It was a geometric distribution. Well, not coincidentally, these two look quite a bit like the other. A geometric distribution has this kind of shape. The exponential distribution has that kind of shape. The geometric is just a discrete version of the exponential. In the Bernoulli case, we are in discrete time. We have a PMF for the time of the first arrival, which is geometric.

In the Poisson case, what we get is the limit of the geometric as you let those lines become closer and closer, which gives you the exponential distribution. Now the Poisson process shares all the memorylessness properties of the Bernoulli process. And the way one can argue is just in terms of this picture.

Since the Poisson process is the limit of Bernoulli processes, whatever qualitative processes you have in the Bernoulli process remain valid for the Poisson process. In particular we have this memorylessness property. You let the Poisson process run for some time, and then you start watching it. What ever happened in the past has no bearing about the future.

Starting from right now, what's going to happen in the future is described again by a Poisson process, in the sense that during every little slot of length delta, there's going to be a probability of lambda delta of having an arrival. And that probably lambda delta is the same-- is always lambda delta-- no matter what happened in the past of the process.

And in particular, we could use this argument to say that the time until the k-th arrival is the time that it takes for the first arrival to happen. OK, let me do it for k equal to two. And then after the first arrival happens, you wait a certain amount of time until the second arrival happens.

Now once the first arrival happened, that's in the past. You start watching. From now on you have mini slots of length delta, each one having a probability of success lambda delta. It's as if we started the Poisson process from scratch. So starting from that time, the time until the next arrival is going to be again an exponential distribution, which doesn't care about what happened in the past, how long it took you for the first arrival.

So these two random variables are going to be independent and exponential, with the same parameter lambda. So among other things, what we have done here is we have essentially derived the PDF of the sum of k independent exponentials. The time of the k-th arrival is the sum of k inter-arrival times. The inter-arrival times are all independent of each other because of memorylessness. And they all have the same exponential distribution.

And by the way, this gives you a way to simulate the Poisson process. If you wanted to simulate it on your computer, you would have one option to break time into tiny, tiny slots. And for every tiny slot, use your random number generator to decide whether there was an arrival or not. To get it very accurate, you would have to use tiny, tiny slots. So that would be a lot of computation.

The more clever way of simulating the Poisson process is you use your random number generator to generate a sample from an exponential distribution and call that your first arrival time. Then go back to the random number generator, generate another independent sample, again from the same exponential distribution. That's the time between the first and the second arrival, and you keep going that way.

So as a sort of a quick summary, this is the big picture. This table doesn't tell you anything new. But it's good to have it as a reference, and to look at it, and to make sure you understand what all the different boxes are. Basically the Bernoulli process runs in discrete time. The Poisson process runs in continuous time.

There's an analogy of arrival rates, p per trial, or intensity per unit time. We did derive, or sketched the derivation for the PMF of the number of arrivals. And the Poisson distribution, which is the distribution that we get, this P_k of t . P_k and t is the limit of the binomial when we take the limit in this particular way, as delta goes to zero, and n goes to infinity.

The geometric becomes an exponential in the limit. And the distribution of the time of the k -th arrival-- we had a closed form formula last time for the Bernoulli process. We got the closed form formula this time for the Poisson process. And we actually used exactly the same argument to get these two closed form formulas.

All right. So now let's talk about adding or merging Poisson processes. And there's two statements that we can make here. One has to do with adding Poisson random variables, just random variables. There's another statement about adding Poisson processes. And the second is a bigger statement than the first. But this is a warm up. Let's work with the first statement.

So the claim is that the sum of independent Poisson random variables is Poisson. OK. So suppose that we have a Poisson process with rate-- just for simplicity-- λ one. And I take the interval from zero to two. And that take then the interval from two until five. The number of arrivals during this interval-- let's call it n from zero to two-- is going to be a Poisson random variable, with parameter, or with mean, two.

The number of arrivals during this interval is n from time two until five. This is again a Poisson random variable with mean equal to three, because the arrival rate is 1 and the duration of the interval is three. These two random variables are independent. They obey the Poisson distribution that we derived before. If you add them, what you get is the number of arrivals during the interval from zero to five.

Now what kind of distribution does this random variable have? Well this is the number of arrivals over an interval of a certain length in a Poisson process. Therefore, this is also Poisson with mean five.

Because for the Poisson process we know that this number of arrivals is Poisson, this is Poisson, but also the number of overall arrivals is also Poisson. This establishes that the sum of a Poisson plus a Poisson random variable gives us another Poisson random variable. So adding Poisson random variables gives us a Poisson random variable.

But now I'm going to make a more general statement that it's not just number of arrivals during a fixed time interval-- it's not just numbers of arrivals for given time intervals-- but rather if you take two different Poisson processes and add them up, the process itself is Poisson in the sense that this process is going to satisfy all the assumptions of a Poisson process.

So the story is that you have a red bulb that flashes at random times at the rate of λ one. It's a Poisson process. You have an independent process where a green bulb flashes at random times. And you happen to be color blind, so you just see when something is flashing. So these two are assumed to be independent Poisson processes. What can we say about the process that you observe?

So in the processes that you observe, if you take a typical time interval of length little delta, what can happen during that little time interval? The red process may have something flashing. So red flashes. Or the red does not. And for the other bulb, the green bulb, there's two possibilities. The green one flashes. And the other possibility is that the green does not.

OK. So there's four possibilities about what can happen during a little slot. The probability that the red one flashes and the green one flashes, what is this probability? It's lambda one delta that the first one flashes, and lambda two delta that the second one does. I'm multiplying probabilities here because I'm making the assumption that the two processes are independent.

OK. Now the probability that the red one flashes is lambda one delta. But the green one doesn't is one, minus lambda two delta. Here the probability would be that the red one does not, times the probability that the green one does. And then here we have the probability that none of them flash, which is whatever is left. But it's one minus lambda one delta, times one minus lambda two delta.

Now we're thinking about delta as small. So think of the case where delta goes to zero, but in a way that we keep the first order terms. We keep the delta terms, but we throw away the delta squared terms. Delta squared terms are much smaller than the delta terms when delta becomes small. If we do that-- if we only keep the order of delta terms-- this term effectively disappears. This is delta squared. So we make it zero. So the probability of having simultaneously a red and a green flash during a little interval is negligible.

What do we get here? Lambda delta times one survives, but this times that doesn't. So we can throw that away. So the approximation that we get is lambda one delta. Similarly here, this goes away. We're left with a lambda two delta. And this is whatever remains, whatever is left.

So what do we have? That there is a probability of seeing a flash, either a red or a green, which is lambda one delta, plus lambda two delta. So if we take a little interval of length delta here, it's going to see an arrival with probability approximately lambda one, plus lambda two, delta.

So every slot in this merged process has an arrival probability with a rate which is the sum of the rates of these two processes. So this is one part of the definition of the Poisson process. There's a few more things that one would need to verify. Namely, that intervals of the same length have the same probability distribution and that different slots are independent of each other.

This can be argued by starting from here because different intervals in this process are independent from each other. Different intervals here are independent from each other. It's not hard to argue that different intervals in the merged process will also be independent of each other.

So the conclusion that comes at the end is that this process is a Poisson process, with a total rate which is equal to the sum of the rate of the two processes. And now if I tell you that an arrival happened in the merged process at a certain time, how likely is it that it came from here? How likely is it?

We go to this picture. Given that an arrival occurred-- which is the event that this or that happened-- what is the probability that it came from the first process, the red one? Well it's the probability of this divided by the probability of this, times that.

Given that this event occurred, you want to find the conditional probability of that sub event. So we're asking the question, out of the total probability of these two, what fraction of that probability is assigned here? And this is lambda one delta, after we ignore the other terms.

This is lambda two delta. So that fraction is going to be lambda one, over lambda one plus lambda two. What does this tell you? If lambda one and lambda two are equal, given that I saw an arrival here, it's equally likely to be red or green. But if the reds have a much higher arrival rate, when I see an arrival here, it's more likely this number will be large. So it's more likely to have come from the red process.

OK so we'll continue with this story and do some applications next time.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

LECTURE 14

The Poisson process

- **Readings:** Start Section 6.2.

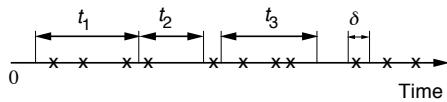
Lecture outline

- Review of Bernoulli process
- Definition of Poisson process
- Distribution of number of arrivals
- Distribution of interarrival times
- Other properties of the Poisson process

Bernoulli review

- Discrete time; success probability p
- Number of arrivals in n time slots: binomial pmf
- Interarrival times: geometric pmf
- Time to k arrivals: Pascal pmf
- Memorylessness

Definition of the Poisson process



- **Time homogeneity:**

$P(k, \tau)$ = Prob. of k arrivals in interval of duration τ

- Numbers of arrivals in disjoint time intervals are **independent**

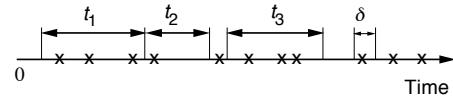
- **Small interval probabilities:**

For VERY small δ :

$$P(k, \delta) \approx \begin{cases} 1 - \lambda\delta, & \text{if } k = 0; \\ \lambda\delta, & \text{if } k = 1; \\ 0, & \text{if } k > 1. \end{cases}$$

– λ : “arrival rate”

PMF of Number of Arrivals N



- Finely discretize $[0, t]$: approximately Bernoulli

- N_t (of discrete approximation): binomial

- Taking $\delta \rightarrow 0$ (or $n \rightarrow \infty$) gives:

$$P(k, \tau) = \frac{(\lambda\tau)^k e^{-\lambda\tau}}{k!}, \quad k = 0, 1, \dots$$

- $E[N_t] = \lambda t, \quad \text{var}(N_t) = \lambda t$

Example

- You get email according to a Poisson process at a rate of $\lambda = 5$ messages per hour. You check your email every thirty minutes.
- Prob(no new messages) =
- Prob(one new message) =

Interarrival Times

- Y_k time of k th arrival

- **Erlang** distribution:

$$f_{Y_k}(y) = \frac{\lambda^k y^{k-1} e^{-\lambda y}}{(k-1)!}, \quad y \geq 0$$

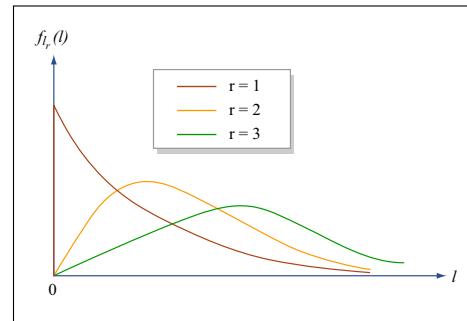


Image by MIT OpenCourseWare.

- Time of first arrival ($k = 1$):
exponential: $f_{Y_1}(y) = \lambda e^{-\lambda y}, \quad y \geq 0$
- **Memoryless** property: The time to the next arrival is independent of the past

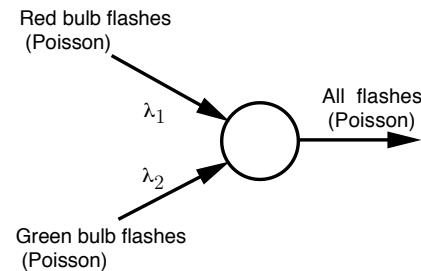
Bernoulli/Poisson Relation



	POISSON	BERNOULLI
Times of Arrival	Continuous	Discrete
Arrival Rate	$\lambda/\text{unit time}$	$p/\text{per trial}$
PMF of # of Arrivals	Poisson	Binomial
Interarrival Time Distr.	Exponential	Geometric
Time to k -th arrival	Erlang	Pascal

Merging Poisson Processes

- Sum of independent Poisson **random variables** is Poisson
- Merging of independent Poisson **processes** is Poisson



- What is the probability that the next arrival comes from the first process?

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2010

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Recitation 15
October 28, 2010

1. Problem 6.14 (a)-(c),(h)-(j), page 330 in text.

Beginning at time $t = 0$, we begin using bulbs, one at a time, to illuminate a room. Bulbs are replaced immediately upon failure. Each new bulb is selected independently by an equally likely choice between a type-A bulb and a type-B bulb. The lifetime, X , of any particular bulb of a particular type is a random variable, independent of everything else, with the following PDF:

$$\text{for type-A Bulbs: } f_X(x) = \begin{cases} e^{-x}, & x \geq 0, \\ 0, & \text{otherwise;} \end{cases}$$
$$\text{for type-B Bulbs: } f_X(x) = \begin{cases} 3e^{-3x}, & x \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

- (a) Find the expected time until the first failure.
 - (b) Find the probability that there are no bulb failures before time t .
 - (c) Given that there are no failures until time t , determine the conditional probability that the first bulb used is a type-A bulb.
 - (d) Determine the probability that the total period of illumination provided by the first two type-B bulbs is longer than that provided by the first type-A bulb.
 - (e) Suppose the process terminates as soon as a total of exactly 12 bulb failures have occurred. Determine the expected value and variance of the total period of illumination provided by type-B bulbs while the process is in operation.
 - (f) Given that there are no failures until time t , find the expected value of the time until the first failure.
2. Problem 6.15 (a)-(c), p. 331 in text.
- A service station handles jobs of two types, A and B. (Multiple jobs can be processed simultaneously.) Arrivals of the two job types are independent Poisson processes with parameters $\lambda_A = 3$ and $\lambda_B = 4$ per minute, respectively. Type A jobs stay in the service station for exactly one minute. Each type B job stays in the service station for a random but integer amount of time which is geometrically distributed, with mean equal to 2, and independent of everything else. The service station started operating at some time in the remote past.
- (a) What is the mean, variance, and PMF of the total number of jobs that arrive within a given three-minute interval?
 - (b) We are told that during a 10-minute interval, exactly 10 new jobs arrived. What is the probability that exactly 3 of them are of type A?
 - (c) At time 0, no job is present in the service station. What is the PMF of the number of type B jobs that arrive in the future, but before the first type A arrival?
3. Let X , Y , and Z be independent exponential random variables with parameters λ , μ , and ν , respectively. Find $\mathbf{P}(X < Y < Z)$.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Recitation 15 Solutions
October 28, 2010

1. (a) Let X be the time until the first bulb failure. Let A (respectively, B) be the event that the first bulb is of type A (respectively, B). Since the two bulb types are equally likely, the total expectation theorem yields

$$\mathbf{E}[X] = \mathbf{E}[X|A]\mathbf{P}(A) + \mathbf{E}[X|B]\mathbf{P}(B) = 1 \cdot \frac{1}{2} + \frac{1}{3} \cdot \frac{1}{2} = \frac{2}{3}.$$

- (b) Let D be the event of no bulb failures before time t . Using the total probability theorem, and the exponential distributions for bulbs of the two types, we obtain

$$\mathbf{P}(D) = \mathbf{P}(D|A)\mathbf{P}(A) + \mathbf{P}(D|B)\mathbf{P}(B) = \frac{1}{2}e^{-t} + \frac{1}{2}e^{-3t}.$$

- (c) We have

$$\mathbf{P}(A|D) = \frac{\mathbf{P}(A \cap D)}{\mathbf{P}(D)} = \frac{\frac{1}{2}e^{-t}}{\frac{1}{2}e^{-t} + \frac{1}{2}e^{-3t}} = \frac{1}{1 + e^{-2t}}.$$

- (d) The lifetime of the first type-A bulb is X_A , with PDF given by:

$$f_{X_A}(x) = \begin{cases} e^{-x} & x \geq 0 \\ 0 & \text{elsewhere} \end{cases}$$

Let Y be the total lifetime of two type-B bulbs. Because the lifetime of each type-B bulb is exponential with $\lambda = 3$, the sum Y has an Erlang distribution of order 2 with $\lambda = 3$. Its PDF is:

$$f_Y(y) = \begin{cases} 9ye^{-3y} & y \geq 0 \\ 0 & \text{elsewhere} \end{cases}$$

$$\begin{aligned} P(G) &= P(Y \geq X_A) \\ &= \int_{-\infty}^{\infty} f_Y(y) \int_{-\infty}^y f_{X_A}(x) dx dy \\ &= \int_0^{\infty} 9ye^{-3y} \int_0^y e^{-x} dx dy = 9 \int_0^{\infty} ye^{-3y} - e^{-x} \Big|_{x=0}^{x=y} dy \\ &= 9 \int_0^{\infty} ye^{-3y} (1 - e^{-y}) dy = 9 \int_0^{\infty} ye^{-3y} - ye^{-4y} dy \\ &= 9 \left(-\frac{1}{3}ye^{-3y} - \frac{1}{9}e^{-3y} + \frac{1}{4}ye^{-4y} + \frac{1}{16}e^{-4y} \right) \Big|_{y=0}^{y=\infty} \\ &= 9 \left(\frac{1}{9} - \frac{1}{16} \right) = \frac{7}{16} \end{aligned}$$

A simpler solution involving no integrals is as follows:

The bulb failure times of interest (1st type- A , 2nd type- B) may be thought of as the arrival

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
 (Fall 2010)

times of two independent Poisson processes of rate $\lambda_A = 1$ and $\lambda_B = 3$. We may imagine that these two processes were split from a joint Poisson process of rate $\lambda_A + \lambda_B$, where the splitting probabilities for each arrival are $P(A) = \frac{\lambda_A}{\lambda_A + \lambda_B} = 1/4$ to process A and $P(B) = \frac{\lambda_B}{\lambda_A + \lambda_B} = 3/4$ to process B. Now we may just focus on whether arrivals to the joint process go to process A or to process B. Each arrival to the joint process corresponds to an independent trial. There are two possible outcomes: the arrival is handed to process A with probability $P(A)$ or the arrival is handed to process B with probability $P(B)$. Then our event of interest occurs when either the first arrival goes to A, or the first arrival goes to B followed by the second going to A. So the corresponding probability is

$$P(A \text{ or } BA) = P(A) + P(BA) = P(A) + P(B)P(A) = 7/16$$

- (e) Let V be the total period of illumination provided by type-B bulbs while the process is in operation. Let N be the number of light bulbs, out of the first 12, that are of type-B. Let X_i be the period of illumination from the i th type-B bulb. We then have $V = Y_1 + \dots + Y_N$. Note that N is a binomial random variable, with parameters $n = 12$ and $p = 1/2$, so that

$$\mathbf{E}[N] = 6, \quad \text{var}(N) = 12 \cdot \frac{1}{2} \cdot \frac{1}{2} = 3.$$

Furthermore, $\mathbf{E}[X_i] = 1/3$ and $\text{var}(X_i) = 1/9$. Using the formulas for the mean and variance of the sum of a random number of random variables, we obtain

$$\mathbf{E}[V] = \mathbf{E}[N]\mathbf{E}[X_i] = 2,$$

and

$$\text{var}(V) = \text{var}(X_i)\mathbf{E}[N] + (\mathbf{E}[X_i])^2\text{var}(N) = \frac{1}{9} \cdot 6 + \frac{1}{9} \cdot 3 = 1$$

- (f) Using the notation in parts (a)-(c), and the result of part (c), we have

$$\begin{aligned} \mathbf{E}[T|D] &= t + \mathbf{E}[T - t|D \cap A]\mathbf{P}(A|D) + \mathbf{E}[T - t|D \cap B]\mathbf{P}(B|D) \\ &= t + 1 \cdot \frac{1}{1 + e^{-2t}} + \frac{1}{3} \left(1 - \frac{1}{1 + e^{-2t}} \right) \\ &= t + \frac{1}{3} + \frac{2}{3} \cdot \frac{1}{1 + e^{-2t}}. \end{aligned}$$

2. (a) The total arrival process corresponds to the merging of two independent Poisson processes, and is therefore Poisson with rate $\lambda = \lambda_A + \lambda_B = 7$. Thus, the number N of jobs that arrive in a given three-minute interval is a Poisson random variable, with $\mathbf{E}[N] = 3\lambda = 21$, $\text{var}(N) = 21$, and PMF

$$p_N(n) = \frac{(21)^n e^{-21}}{n!}, \quad n = 0, 1, 2, \dots$$

- (b) Each of these 10 jobs has probability $\lambda_A/(\lambda_A + \lambda_B) = 3/7$ of being type A, independently of the others. Thus, the binomial PMF applies and the desired probability is equal to

$$\binom{10}{3} \left(\frac{3}{7}\right)^3 \left(\frac{4}{7}\right)^7$$

- (c) Each future arrival is of type A with probability $\lambda_A/(\lambda_A + \lambda_B) = 3/7$ of being type A, independently of the others. Thus, the number K of arrivals until the first type A arrival is geometric with parameter $3/7$. The number of type B arrivals before the first type A arrival is equal to $K - 1$, and its PMF is similar to a geometric, except that it is shifted by one unit to the left. In particular,

$$p_K(k) = \left(\frac{3}{7}\right) \left(\frac{4}{7}\right)^k, \quad k = 0, 1, 2, \dots$$

3. The event $\{X < Y < Z\}$ can be expressed as $\{X < \min\{Y, Z\}\} \cap \{Y < Z\}$. Let Y and Z be the 1st arrival times of two independent Poisson processes with rates μ and ν . By merging the two processes, it should be clear that $Y < Z$ if and only if the first arrival of the merged process comes from the original process with rate μ , and thus

$$\mathbf{P}(Y < Z) = \frac{\mu}{\mu + \nu}.$$

Let X be the 1st arrival time of a third independent Poisson process with rate λ . Now $\{X < \min\{Y, Z\}\}$ if and only if the first arrival of the Poisson process obtained by merging the two processes with rates λ and $\mu + \nu$ comes from the original process with rate λ , and thus

$$\mathbf{P}(X < \min\{Y, Z\}) = \frac{\lambda}{\lambda + \mu + \nu}.$$

Note that the event $\{X < \min\{Y, Z\}\}$ is independent of the event $\{Y < Z\}$, as the time of the first arrival of the merged process with rate $\mu + \nu$ is independent of whether that first arrival comes from the process with rate μ or the process with rate ν . Hence,

$$\begin{aligned} \mathbf{P}(X < Y < Z) &= \mathbf{P}(X < \min\{Y, Z\}) \cdot \mathbf{P}(Y < Z) \\ &= \frac{\lambda\mu}{(\lambda + \mu + \nu)(\mu + \nu)}. \end{aligned}$$

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Tutorial 7
October 28/29, 2010

1. Alice and Bob alternate playing at the casino table. (Alice starts and plays at odd times $i = 1, 3, \dots$; Bob plays at even times $i = 2, 4, \dots$) At each time i , the net gain of whoever is playing is a random variable G_i with the following PMF:

$$p_G(g) = \begin{cases} \frac{1}{3} & g = -2, \\ \frac{1}{3} & g = 1, \\ \frac{1}{6} & g = 3, \\ 0 & \text{otherwise} \end{cases}$$

Assume that the net gains at different times are independent. We refer to an outcome of -2 as a “loss.”

- (a) They keep gambling until the first time where a loss by Bob immediately follows a loss by Alice. Write down the PMF of the total number of rounds played. (A round consists of two plays, one by Alice and then one by Bob.)
- (b) Write down the PMF for Z , defined as the time at which Bob has his third loss.
- (c) Let N be the number of rounds until each one of them has won at least once. Find $\mathbf{E}[N]$.

2. Problem 6.6, page 328 in text.

Sum of a geometric number of independent geometric random variables

Let $Y = X_1 + \dots + X_N$, where the random variable X_i are geometric with parameter p , and N is geometric with parameter q . Assume that the random variables N, X_1, X_2, \dots are independent. Show that Y is geometric with parameter pq . *Hint:* Interpret the various random variables in terms of a split Bernoulli process.

3. A train bridge is constructed across a wide river. Trains arrive at the bridge according to a Poisson process of rate $\lambda = 3$ per day.

- If a train arrives on day 0, find the probability that there will be no trains on days 1, 2, and 3.
- Find the probability that the next train to arrive after the first train on day 0, takes more than 3 days to arrive.
- Find the probability that no trains arrive in the first 2 days, but 4 trains arrive on the 4th day.
- Find the probability that it takes more than 2 days for the 5th train to arrive at the bridge.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Tutorial 7: Solutions

1. (a) For each round, the probability that both Alice and Bob have a loss is $\frac{1}{3} \cdot \frac{1}{3} = \frac{1}{9}$. Let random variable X represent the total number of rounds played until the first time where they both have a loss. Then X is a geometric random variable with parameter $p = 1/9$ and has the following PMF.

$$p_X(x) = (1 - p)^{x-1} p = \left(\frac{8}{9}\right)^{x-1} \left(\frac{1}{9}\right), \quad x = 1, 2, \dots$$

- (b) First, consider the number of games, K_3 Bob played until his third loss. Random variable K_3 is a Pascal random variable and has the following PMF.

$$p_{K_3}(k) = \binom{k-1}{3-1} \left(\frac{1}{3}\right)^3 \left(\frac{2}{3}\right)^{k-3} \quad k = 3, 4, 5, \dots$$

In this question, we are interested in another random variable Z defined as the time at which Bob has his third loss. Note that $Z = 2K_3$. By changing variables, we obtain

$$p_Z(z) = \binom{\frac{z}{2}-1}{3-1} \left(\frac{1}{3}\right)^3 \left(\frac{2}{3}\right)^{\frac{z}{2}-3} \quad z = 6, 8, 10, \dots$$

- (c) Let A be the event that Alice wins, and Let B be the event that Bob wins. The event $A \cup B$ is then the event that either A wins or B wins or both A and B win, and the event $A \cap B$ is the event that both A and B win. Suppose we observe this gambling process, and let U be a random variable indicating the number of rounds we see until at least one of them wins. Random variable U is a geometric random variable with parameter $p = P(A \cup B) = 1 - \frac{1}{3} \cdot \frac{1}{3}$.

Consider another random variable V representing the number of additional rounds we have to observe until the other wins. If both Alice and Bob win at the U th round, then $V = 0$. This occurs with probability $P(A \cap B | A \cup B) = \frac{2}{9}$. If Alice wins the U th round, then the time V until Bob wins is a geometric random variable with parameter $p = 1/2 + 1/6 = 2/3$. This occurs with probability $P(A \cap B^c | A \cup B) = \frac{1}{3} \cdot \frac{2}{3}$. Likewise, if Bob wins the U th round, then the time V until Alice wins is a geometric random variable with parameter $p = 1/2 + 1/6 = 2/3$. This occurs with probability $P(B \cap A^c | A \cup B) = \frac{1}{3} \cdot \frac{2}{3}$. The number of rounds until each one of them has won at least once, N is

$$N = U + V$$

The expectation of N is then:

$$\begin{aligned} E[N] &= E[U] + E[V] \\ &= \frac{1}{\frac{8}{9}} + 0 \cdot P(A \cap B | A \cup B) + \frac{1}{\frac{2}{3}} P(A | A \cup B) + \frac{1}{\frac{2}{3}} P(B | A \cup B) \\ &= 9/8 + \frac{3}{2} \frac{\frac{1}{2} \frac{2}{3}}{\frac{8}{9}} + \frac{3}{2} \frac{\frac{1}{2} \frac{2}{3}}{\frac{8}{9}} \\ &= 15/8 \end{aligned}$$

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
 (Fall 2010)

There is another approach to this problem. Consider the following partition.

- A_1 : both win first round
- A_2 : Only Alice wins first round
- A_3 : Only Bob wins first round
- A_4 : both lose first round

Event A_1 occurs with probability $\frac{2}{3} \cdot \frac{2}{3}$. Event A_2 occurs with probability $\frac{2}{3} \cdot \frac{1}{3}$. Event A_3 occurs with probability $\frac{1}{3} \cdot \frac{2}{3}$. Event A_4 occurs with probability $\frac{1}{3} \cdot \frac{1}{3}$. When event A_2 (A_3) occurs, the distribution on the time until Bob (Alice) wins is a geometric random variable with mean $\frac{1}{\frac{2}{3}} = \frac{3}{2}$. When event A_4 occurs, the additional time until Alice and Bob win is distributed identically to that at time 0 by the fresh-start property. By the total expectation theorem,

$$\begin{aligned} E[N] &= E[N|A_1]P(A_1) + E[N|A_2]P(A_2) + E[N|A_3]P(A_3) + E[N|A_4]P(A_4) \\ &= 1 \cdot \left(\frac{2}{3} \cdot \frac{2}{3}\right) + \left(1 + \frac{1}{\frac{2}{3}}\right) \cdot \left(\frac{1}{3} \cdot \frac{2}{3}\right) + \left(\frac{1}{3} \cdot \frac{2}{3}\right) + (1 + E[N]) \cdot \left(\frac{1}{3} \cdot \frac{1}{3}\right) \end{aligned}$$

Solving for $E[N]$, we get $E[N] = \frac{15}{8}$.

2. Problem 6.6, page 328 in text. See text for solutions.
3. (a) The number of trains arriving on days 1, 2, and 3 is independent of the number of trains arriving on day 0. Let N denote the total number of trains that arrive on days 1, 2, and 3. Then N is a Poisson random variable with parameter $3\lambda = 9$, and we have

$$\begin{aligned} P(\text{no train on days 1,2,3} \mid \text{one train on day 1}) &= P(\text{no train on days 1,2,3}) \\ &= P(N = 0) \\ &= \frac{e^{-9}9^0}{0!} \\ &= e^{-9}. \end{aligned}$$

- (b) The event that the next arrival is more than three days after the train arrival on day 0 is the same as the event that there are zero arrivals in the three days after the train arrival on day 0. Therefore the required probability is the same as that found in part (a), namely, e^{-9} .
- (c) The number of trains arriving in the first 2 days is independent of the number of trains arriving on day 4. Therefore, we have

$$\begin{aligned} P(\text{0 trains in first 2 days and 4 trains on day 4}) &= P(\text{0 trains in first 2 days}) \cdot P(\text{4 trains on day 4}) \\ &= e^{-2\lambda} \frac{(2\lambda)^0}{0!} \cdot e^{-\lambda} \frac{\lambda^4}{4!} \\ &= e^{-9} \frac{3^4}{4!}. \end{aligned}$$

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Fall 2010)

- (d) The event that it takes more than 2 days for the 5th arrival is equivalent to the event that there are at most 4 arrivals in the first 2 days. Therefore the required probability is equal to

$$\begin{aligned}\sum_{k=0}^4 P(\text{exactly } k \text{ arrivals in first 2 days}) &= \sum_{k=0}^4 e^{-2\lambda} \frac{(2\lambda)^k}{k!} \\&= e^{-2\lambda} \left(\frac{(2\lambda)^0}{0!} + \frac{(2\lambda)^1}{1!} + \frac{(2\lambda)^2}{2!} + \frac{(2\lambda)^3}{3!} + \frac{(2\lambda)^4}{4!} \right) \\&= e^{-6}(1 + 6 + 18 + 36 + 54) \\&= 115e^{-6}.\end{aligned}$$

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013

Transcript – Lecture 15

The following content is provided under a Creative Commons license. Your support will help MIT OpenCourseWare continue to offer high quality educational resources for free. To make a donation or view additional materials from hundreds of MIT courses, visit MIT OpenCourseWare at ocw.mit.edu.

JOHN TSITSIKLIS: Today we're going to finish our discussion of the Poisson process. We're going to see a few of its properties, do a few interesting problems, some more interesting than others. So go through a few examples and then we're going to talk about some quite strange things that happen with the Poisson process.

So the first thing is to remember what the Poisson processes is. It's a model, let's say, of arrivals of customers that are, in some sense, quote unquote, completely random, that is a customer can arrive at any point in time. All points in time are equally likely. And different points in time are sort of independent of other points in time. So the fact that I got an arrival now doesn't tell me anything about whether there's going to be an arrival at some other time.

In some sense, it's a continuous time version of the Bernoulli process. So the best way to think about the Poisson process is that we divide time into extremely tiny slots. And in each time slot, there's an independent possibility of having an arrival. Different time slots are independent of each other. On the other hand, when the slot is tiny, the probability for obtaining an arrival during that tiny slot is itself going to be tiny.

So we capture these properties into a formal definition what the Poisson process is. We have a probability mass function for the number of arrivals, k , during an interval of a given length. So this is the sort of basic description of the distribution of the number of arrivals. So τ is fixed. And k is the parameter. So when we add over all k 's, the sum of these probabilities has to be equal to 1.

There's a time homogeneity assumption, which is hidden in this, namely, the only thing that matters is the duration of the time interval, not where the time interval sits on the real axis. Then we have an independence assumption. Intervals that are disjoint are statistically independent from each other. So any information you give me about arrivals during this time interval doesn't change my beliefs about what's going to happen during another time interval. So this is a generalization of the idea that we had in Bernoulli processes that different time slots are independent of each other.

And then to specify this function, the distribution of the number of arrivals, we sort of go in stages. We first specify this function for the case where the time interval is very small. And I'm telling you what those probabilities will be. And based on these then, we do some calculations and to find the formula for the distribution of the number of arrivals for intervals of a general duration. So for a small duration, δ , the probability of obtaining 1 arrival is $\lambda\delta$. The remaining probability is assigned to the event that we get to no arrivals during that interval.

The probability of obtaining more than 1 arrival in a tiny interval is essentially 0. And when we say essentially, it's means modular, terms that of order delta squared. And when delta is very small, anything which is delta squared can be ignored. So up to delta squared terms, that's what happened during a little interval.

Now if we know the probability distribution for the number of arrivals in a little interval. We can use this to get the distribution for the number of arrivals over several intervals. How do we do that? The big interval is composed of many little intervals. Each little interval is independent from any other little interval, so is it is as if we have a sequence of Bernoulli trials. Each Bernoulli trial is associated with a little interval and has a small probability of obtaining a success or an arrival during that mini-slot.

On the other hand, when delta is small, and you take a big interval and chop it up, you get a large number of little intervals. So what we essentially have here is a Bernoulli process, in which is the number of trials is huge but the probability of success during any given trial is tiny. The average number of trials ends up being proportional to the length of the interval. If you have twice as large an interval, it's as if you're having twice as many over these mini-trials, so the expected number of arrivals will increase proportionately.

There's also this parameter lambda, which we interpret as expected number of arrivals per unit time. And it comes in those probabilities here. When you double lambda, this means that a little interval is twice as likely to get an arrival. So you would expect to get twice as many arrivals as well. That's why the expected number of arrivals during an interval of length tau also scales proportional to this parameter lambda. Somewhat unexpectedly, it turns out that the variance of the number of arrivals is also the same as the mean. This is a peculiarity that happens in the Poisson process.

So this is one way of thinking about Poisson process, in terms of little intervals, each one of which has a tiny probability of success. And we think of the distribution associated with that process as being described by this particular PMF. So this is the PMF for the number of arrivals during an interval of a fixed duration, tau. It's a PMF that extends all over the entire range of non-negative integers.

So the number of arrivals you can get during an interval for certain length can be anything. You can get as many arrivals as you want. Of course the probability of getting a zillion arrivals is going to be tiny. But in principle, this is possible. And that's because an interval, even if it's a fixed length, consists of an infinite number of mini-slots in some sense. You can divide, chop it up, into as many mini-slots as you want. So in principle, it's possible that every mini-slot gets an arrival. In principle, it's possible to get an arbitrarily large number of arrivals.

So this particular formula here is not very intuitive when you look at it. But it's a legitimate PMF. And it's called the Poisson PMF. It's the PMF that describes the number of arrivals. So that's one way of thinking about the Poisson process, where the basic object of interest would be this PMF and you try to work with it.

There's another way of thinking about what happens in the Poisson process. And this has to do with letting things evolve in time. You start at time 0. There's going to be a time at which the first arrival occurs, and call that time T_1 . This time turns out to have an exponential distribution with parameter λ . Once you get an arrival, it's as if the process starts fresh.

The best way to understand why this is the case is by thinking in terms of the analogy with the Bernoulli process. If you believe that statement for the Bernoulli process, since this is a limiting case, it should also be true. So starting from this time, we're going to wait a random amount of time until we get the second arrival. This random amount of time, let's call it T_2 . This time, T_2 is also going to have an exponential distribution with the same parameter, λ . And these two are going to be independent of each other. OK?

So the Poisson process has all the same memorylessness properties that the Bernoulli process has. What's another way of thinking of this property? So think of a process where you have a light bulb. The time at the light bulb burns out, you can model it by an exponential random variable. And suppose that they tell you that so far, we're sitting at some time, T . And I tell you that the light bulb has not yet burned out. What does this tell you about the future of the light bulb? Is the fact that they didn't burn out, so far, is it good news or is it bad news? Would you rather keep this light bulb that has worked for t time steps and is still OK? Or would you rather use a new light bulb that starts new at that point in time?

Because of the memorylessness property, the past of that light bulb doesn't matter. So the future of this light bulb is statistically the same as the future of a new light bulb. For both of them, the time until they burn out is going to be described by an exponential distribution. So one way that people described the situation is to say that used is exactly as good as a new. So a used one is no worse than a new one. A used one is no better than a new one. So a used light bulb that hasn't yet burnt out is exactly as good as a new light bulb. So that's another way of thinking about the memorylessness that we have in the Poisson process.

Back to this picture. The time until the second arrival is the sum of two independent exponential random variables. So, in principle, you can use the convolution formula to find the distribution of T_1 plus T_2 , and that would be what we call Y_2 , the time until the second arrival. But there's also a direct way of obtaining the distribution of Y_2 , and this is the calculation that we did last time on the blackboard. And actually, we did it more generally. We found the time until the k -th arrival occurs. It has a closed form formula, which is called the Erlang distribution with k degrees of freedom.

So let's see what's going on here. It's a distribution of what kind? It's a continuous distribution. It's a probability density function. This is because the time is a continuous random variable. Time is continuous. Arrivals can happen at any time. So we're talking about the PDF. This k is just the parameter of the distribution. We're talking about the k -th arrival, so k is a fixed number. λ is another parameter of the distribution, which is the arrival rate. So it's a PDF over the Y 's, whereas λ and k are parameters of the distribution. OK.

So this was what we knew from last time. Just to get some practice, let us do a problem that's not too difficult, but just to see how we use the various formulas that we have. So Poisson was a

mathematician, but Poisson also means fish in French. So Poisson goes fishing. And let's assume that fish are caught according to a Poisson process.

That's not too bad an assumption. At any given point in time, you have a little probability that a fish would be caught. And whether you catch one now is sort of independent about whether at some later time a fish will be caught or not. So let's just make this assumption. And suppose that the rules of the game are that you-- Fish are being called it the certain rate of 0.6 per hour. You fish for 2 hours, no matter what. And then there are two possibilities. If I have caught a fish, I stop and go home. So if some fish have been caught, so there's at least 1 arrival during this interval, I go home. Or if nothing has been caught, I continue fishing until I catch something. And then I go home. So that's the description of what is going to happen.

And now let's starts asking questions of all sorts. What is the probability that I'm going to be fishing for more than 2 hours? I will be fishing for more than 2 hours, if and only if no fish were caught during those 2 hours, in which case, I will have to continue. Therefore, this is just this quantity. The probability of catching 2 fish in-- of catching 0 fish in the next 2 hours, and according to the formula that we have, this is going to be e to the minus lambda times how much time we have.

There's another way of thinking about this. The probability that I fish for more than 2 hours is the probability that the first catch happens after time 2, which would be the integral from 2 to infinity of the density of the first arrival time. And that density is an exponential. So you do the integral of an exponential, and, of course, you would get the same answer. OK. That's easy.

So what's the probability of fishing for more than 2 but less than 5 hours? What does it take for this to happen? For this to happen, we need to catch 0 fish from time 0 to 2 and catch the first fish sometime between 2 and 5. So if you-- one way of thinking about what's happening here might be to say that there's a Poisson process that keeps going on forever. But as soon as I catch the first fish, instead of continuing fishing and obtaining those other fish I just go home right now.

Now the fact that I go home before time 5 means that, if I were to stay until time 5, I would have caught at least 1 fish. I might have caught more than 1. So the event of interest here is that the first catch happens between times 2 and 5. So one way of calculating this quantity would be-- Its the probability that the first catch happens between times 2 and 5. Another way to deal with it is to say, this is the probability that I caught 0 fish in the first 2 hours and then the probability that I catch at least 1 fish during the next 3 hours.

This. What is this? The probability of 0 fish in the next 3 hours is the probability of 0 fish during this time. 1 minus this is the probability of catching at least 1 fish, of having at least 1 arrival, between times 2 and 5. If there's at least 1 arrival between times 2 and 5, then I would have gone home by time 5. So both of these, if you plug-in numbers and all that, of course, are going to give you the same answer.

Now next, what's the probability that I catch at least 2 fish? In which scenario are we? Under this scenario, I go home when I catch my first fish. So in order to catch at least 2 fish, it must be in

this case. So this is the same as the event that I catch at least 2 fish during the first 2 time steps. So it's going to be the probability from 2 to infinity, the probability that I catch 2 fish, or that I catch 3 fish, or I catch more than that.

So it's this quantity. k is the number of fish that I catch. At least 2, so k goes from 2 to infinity. These are the probabilities of catching a number k of fish during this interval. And if you want a simpler form without an infinite sum, this would be 1 minus the probability of catching 0 fish, minus the probability of catching 1 fish, during a time interval of length 2. Another way to think of it. I'm going to catch 2 fish, at least 2 fish, if and only if the second fish caught in this process happens before time 2. So that's another way of thinking about the same event. So it's going to be the probability that the random variable Y_2 , the arrival time over the second fish, is less than or equal to 2. OK.

The next one is a little trickier. Here we need to do a little bit of divide and conquer. Overall, in this expedition, what the expected number of fish to be caught? One way to think about it is to try to use the total expectations theorem. And think of expected number of fish, given this scenario, or expected number of fish, given this scenario. That's a little more complicated than the way I'm going to do it.

The way I'm going to do is to think as follows-- Expected number of fish is the expected number of fish caught between times 0 and 2 plus expected number of fish caught after time 2. So what's the expected number caught between time 0 and 2? This is λt . So λ is 0.6 times 2. This is the expected number of fish that are caught between times 0 and 2.

Now let's think about the expected number of fish caught afterwards. How many fish are being caught afterwards? Well it depends on the scenario. If we're in this scenario, we've gone home and we catch 0. If we're in this scenario, then we continue fishing until we catch one. So the expected number of fish to be caught after time 2 is going to be the probability of this scenario times 1. And the probability of that scenario is the probability that they call it's 0 fish during the first 2 time steps times 1, which is the number of fish I'm going to catch if I continue.

The expected total fishing time we can calculate exactly the same way. I'm jumping to the last one. My total fishing time has a period of 2 time steps. I'm going to fish for 2 time steps no matter what. And then if I caught 0 fish, which happens with this probability, my expected time is going to be the expected time from here onwards, which is the expected value of this geometric random variable with parameter λ . So the expected time is 1 over λ . And in our case this, is $1/0.6$.

Finally, if I tell you that I have been fishing for 4 hours and nothing has been caught so far, how much do you expect this quantity to be? Here is the story that, again, that for the Poisson process used is as good as new. The process does not have any memory. Given what happens in the past doesn't matter for the future. It's as if the process starts new at this point in time. So this one is going to be, again, the same exponentially distributed random variable with the same parameter λ .

So expected time until an arrival comes is an exponential distribution -- has an exponential distribution with parameter lambda, no matter what has happened in the past. Starting from now and looking into the future, it's as if the process has just started. So it's going to be 1 over lambda, which is 1/0.6. OK.

Now our next example is going to be a little more complicated or subtle. But before we get to the example, let's refresh our memory about what we discussed last time about merging Poisson independent Poisson processes. Instead of drawing the picture that way, another way we could draw it could be this. We have a Poisson process with rate lambda1, and a Poisson process with rate lambda2. They have, each one of these, have their arrivals. And then we form the merged process. And the merged process records an arrival whenever there's an arrival in either of the two processes.

This process in that process are assumed to be independent of each other. Now different times in this process and that process are independent of each other. So what happens in these two time intervals is independent from what happens in these two time intervals. These two time intervals determine what happens here. These two time intervals determine what happens there. So because these are independent from these, this means that this is also independent from that. So the independence assumption is satisfied for the merged process.

And the merged process turns out to be a Poisson process. And if you want to find the arrival rate for that process, you argue as follows. During a little interval of length delta, we have probability lambda1 delta of having an arrival in this process. We have probability lambda2 delta of an arrival in this process, plus second order terms in delta, which we're ignoring. And then you do the calculation and you find that in this process, you're going to have an arrival probability, which is lambda1 plus lambda2, again ignoring second order in delta-- terms that are second order in delta. So the merged process is a Poisson process whose arrival rate is the sum of the arrival rates of the individual processes.

And the calculation we did at the end of the last lecture-- If I tell you that the new arrival happened here, where did that arrival come from? Did it come from here or from there? If the lambda1 is equal to lambda2, then by symmetry you would say that it's equally likely to have come from here or to come from there. But if this lambda is much bigger than that lambda, the fact that they saw an arrival is more likely to have come from there. And the formula that captures this is the following. This is the probability that my arrival has come from this particular stream rather than that particular stream.

So when an arrival comes and you ask, what is the origin of that arrival? It's as if I'm flipping a coin with these odds. And depending on outcome of that coin, I'm going to tell you came from there or it came from there. So the origin of an arrival is either this stream or that stream. And this is the probability that the origin of the arrival is that one. Now if we look at 2 different arrivals, and we ask about their origins-- So let's think about the origin of this arrival and compare it with the origin that arrival.

The origin of this arrival is random. It could be right be either this or that. And this is the relevant probability. The origin of that arrival is random. It could be either here or is there, and again,

with the same relevant probability. Question. The origin of this arrival, is it dependent or independent from the origin that arrival? And here's how the argument goes. Separate times are independent. Whatever has happened in the process during this set of times is independent from whatever happened in the process during that set of times. Because different times have nothing to do with each other, the origin of this, of an arrival here, has nothing to do with the origin of an arrival there. So the origins of different arrivals are also independent random variables.

So if I tell you that-- yeah. OK. So it as if that each time that you have an arrival in the merge process, it's as if you're flipping a coin to determine where did that arrival came from and these coins are independent of each other. OK. OK.

Now we're going to use this-- what we know about merged processes to solve the problem that would be harder to do, if you were not using ideas from Poisson processes. So the formulation of the problem has nothing to do with the Poisson process. The formulation is the following. We have 3 light-bulbs. And each light bulb is independent and is going to die out at the time that's exponentially distributed. So 3 light bulbs. They start their lives and then at some point they die or burn out. So let's think of this as X, this as Y, and this as Z.

And we're interested in the time until the last light-bulb burns out. So we're interested in the maximum of the 3 random variables, X, Y, and Z. And in particular, we want to find the expected value of this maximum. OK.

So you can do derived distribution, use the expected value rule, anything you want. You can get this answer using the tools that you already have in your hands. But now let us see how we can connect to this picture with a Poisson picture and come up with the answer in a very simple way. What is an exponential random variable? An exponential random variable is the first act in the long play that involves a whole Poisson process. So an exponential random variable is the first act of a Poisson movie. Same thing here. You can think of this random variable as being part of some Poisson process that has been running. So it's part of this bigger picture.

We're still interested in the maximum of the 3. The other arrivals are not going to affect our answers. It's just, conceptually speaking, we can think of the exponential random variable as being embedded in a bigger Poisson picture. So we have 3 Poisson process that are running in parallel. Let us split the expected time until the last burnout into pieces, which is time until the first burnout, time from the first until the second, and time from the second until the third. And find the expected values of each one of these pieces.

What can we say about the expected value of this? This is the first arrival out of all of these 3 Poisson processes. It's the first event that happens when you look at all of these processes simultaneously. So 3 Poisson processes running in parallel. We're interested in the time until one of them, any one of them, gets in arrival. Rephrase. We merged the 3 Poisson processes, and we ask for the time until we observe an arrival in the merged process.

When 1 of the 3 gets an arrival for the first time, the merged process gets its first arrival. So what's the expected value of this time until the first burnout? It's going to be the expected value of a Poisson random variable. So the first burnout is going to have an expected value, which is--

OK. It's a Poisson process. The merged process of the 3 has a collective arrival rate, which is 3 times lambda.

So this is the parameter over the exponential distribution that describes the time until the first arrival in the merged process. And the expected value of this random variable is 1 over that. When you have an exponential random variable with parameter lambda, the expected value of that random variable is 1 over lambda. Here we're talking about the first arrival time in a process with rate 3 lambda. The expected time until the first arrival is 1 over (3 lambda). Alright.

So at this time, this bulb, this arrival happened, this bulb has been burned. So we don't care about that bulb anymore. We start at this time, and we look forward. This bulb has been burned. So let's just look forward from now on. What have we got? We have two bulbs that are burning. We have a Poisson process that's the bigger picture of what could happen to that light bulb, if we were to keep replacing it. Another Poisson process. These two processes are, again, independent.

From this time until that time, how long does it take? It's the time until either this process records an arrival or that process records an arrival. That's the same as the time that the merged process of these two records an arrival. So we're talking about the expected time until the first arrival in a merged process. The merged process is Poisson. It's Poisson with rate 2 lambda. So that extra time is going to take-- the expected value is going to be 1 over the (rate of that Poisson process). So 1 over (2 lambda) is the expected value of this random variable.

So at this point, this bulb now is also burned. So we start looking from this time on. That part of the picture disappears. Starting from this time, what's the expected value until that remaining light-bulb burns out? Well, as we said before, in a Poisson process or with exponential random variables, we have memorylessness. A used bulb is as good as a new one. So it's as if we're starting from scratch here. So this is going to be an exponential random variable with parameter lambda. And the expected value of it is going to be 1 over lambda.

So the beauty of approaching this problem in this particular way is, of course, that we manage to do everything without any calculus at all, without striking an integral, without trying to calculate expectations in any form. Most of the non-trivial problems that you encounter in the Poisson world basically involve tricks of these kind. You have a question and you try to rephrase it, trying to think in terms of what might happen in the Poisson setting, use memorylessness, use merging, et cetera, et cetera.

Now we talked about merging. It turns out that the splitting of Poisson processes also works in a nice way. The story here is exactly the same as for the Bernoulli process. So I'm having a Poisson process. And each time, with some rate lambda, and each time that an arrival comes, I'm going to send it to that stream and the record an arrival here with some probability P. And I'm going to send it to the other stream with some probability 1 minus P. So either of this will happen or that will happen, depending on the outcome of the coin flip that I do. Each time that then arrival occurs, I flip a coin and I decide whether to record it here or there. This is called splitting a Poisson process into two pieces.

What kind of process do we get here? If you look at the little interval for length delta, what's the probability that this little interval gets an arrival? It's the probability that this one gets an arrival, which is lambda delta times the probability that after I get an arrival my coin flip came out to be that way, so that it sends me there. So this means that this little interval is going to have probability lambda delta P. Or maybe more suggestively, I should write it as lambda P times delta.

So every little interval has a probability of an arrival proportional to delta. The proportionality factor is lambda P. So lambda P is the rate of that process. And then you go through the mental exercise that you went through for the Bernoulli process to argue that a different intervals here are independent and so on. And that completes checking that this process is going to be a Poisson process.

So when you split a Poisson process by doing independent coin flips each time that something happens, the processes that you get is again a Poisson process, but of course with a reduced rate. So instead of the word splitting, sometimes people also use the words thinning-out. That is, out of the arrivals that came, you keep a few but throw away a few. OK.

So now the last topic over this lecture is a quite curious phenomenon that goes under the name of random incidents. So here's the story. Buses have been running on Mass Ave. from time immemorial. And the bus company that runs the buses claims that they come as a Poisson process with some rate, let's say, of 4 buses per hour. So that the expected time between bus arrivals is going to be 15 minutes. OK. Alright.

So people have been complaining that they have been showing up there. They think the buses are taking too long. So you are asked to investigate. Is the company-- Does it operate according to its promises or not. So you send an undercover agent to go and check the interarrival times of the buses. Are they 15 minutes? Or are they longer?

So you put your dark glasses and you show up at the bus stop at some random time. And you go and ask the guy in the falafel truck, how long has it been since the last arrival? So of course that guy works for the FBI, right? So they tell you, well, it's been, let's say, 12 minutes since the last bus arrival. And then you say, "Oh, 12 minutes. Average time is 15. So a bus should be coming any time now."

Is that correct? No, you wouldn't think that way. It's a Poisson process. It doesn't matter how long it has been since the last bus arrival. So you don't go through that fallacy. Instead of predicting how long it's going to be, you just sit down there and wait and measure the time. And you find that this is, let's say, 11 minutes. And you go to your boss and report, "Well, it took-- I went there and the time from the previous bus to the next one was 23 minutes. It's more than the 15 that they said."

So go and do that again. You go day after day. You keep these statistics of the length of this interval. And you tell your boss it's a lot more than 15. It tends to be more like 30 or so. So the bus company is cheating us. Does the bus company really run Poisson buses at the rate that they

have promised? Well let's analyze the situation here and figure out what the length of this interval should be, on the average.

The naive argument is that this interval is an interarrival time. And interarrival times, on the average, are 15 minutes, if the company runs indeed Poisson processes with these interarrival times. But actually the situation is a little more subtle because this is not a typical interarrival interval. This interarrival interval consists of two pieces. Let's call them T_1 and T_1 prime. What can you tell me about those two random variables? What kind of random variable is T_1 ? Starting from this time, with the Poisson process, the past doesn't matter. It's the time until an arrival happens. So T_1 is going to be an exponential random variable with parameter lambda.

So in particular, the expected value of T_1 is going to be 15 by itself. How about the random variable T_1 prime. What kind of random variable is it? This is like the first arrival in a Poisson process that runs backwards in time. What kind of process is a Poisson process running backwards in time? Let's think of coin flips. Suppose you have a movie of coin flips. And for some accident, that fascinating movie, you happen to watch it backwards. Will it look any different statistically? No. It's going to be just the sequence of random coin flips.

So a Bernoulli process that's runs in reverse time is statistically identical to a Bernoulli process in forward time. The Poisson process is a limit of the Bernoulli. So, same story with the Poisson process. If you run it backwards in time it looks the same. So looking backwards in time, this is a Poisson process. And T_1 prime is the time until the first arrival in this backward process.

So T_1 prime is also going to be an exponential random variable with the same parameter, lambda. And the expected value of T_1 prime is 15. Conclusion is that the expected length of this interval is going to be 30 minutes. And the fact that this agent found the average to be something like 30 does not contradict the claims of the bus company that they're running Poisson buses with a rate of lambda equal to 4. OK.

So maybe the company can this way-- they can defend themselves in court. But there's something puzzling here. How long is the interarrival time? Is it 15? Or is it 30? On the average. The issue is what do we mean by a typical interarrival time. When we say typical, we mean some kind of average. But average over what? And here's two different ways of thinking about averages. You number the buses. And you have bus number 100. You have bus number 101, bus number 102, bus number 110, and so on.

One way of thinking about averages is that you pick a bus number at random. I pick, let's say, that bus, all buses being sort of equally likely to be picked. And I measure this interarrival time. So for a typical bus. Then, starting from here until there, the expected time has to be 1 over lambda, for the Poisson process.

But what we did in this experiment was something different. We didn't pick a bus at random. We picked a time at random. And if the picture is, let's say, this way, I'm much more likely to pick this interval and therefore this interarrival time, rather than that interval. Because, this interval corresponds to very few times. So if I'm picking a time at random and, in some sense, let's say,

uniform, so that all times are equally likely, I'm much more likely to fall inside a big interval rather than a small interval.

So a person who shows up at the bus stop at a random time. They're selecting an interval in a biased way, with the bias favor of longer intervals. And that's why what they observe is a random variable that has a larger expected value than the ordinary expected value.

So the subtlety here is to realize that we're talking between two different kinds of experiments. Picking a bus number at random verses picking an interval at random with a bias in favor of longer intervals. Lots of paradoxes that one can cook up using Poisson processes and random processes in general often have to do with the story of this kind.

The phenomenon that we had in this particular example also shows up in general, whenever you have other kinds of arrival processes. So the Poisson process is the simplest arrival process there is, where the interarrival times are exponential random variables. There's a larger class of models. They're called renewal processes, in which, again, we have a sequence of successive arrivals, interarrival times are identically distributed and independent, but they may come from a general distribution.

So to make the same point of the previous example but in a much simpler setting, suppose that bus interarrival times are either 5 or 10 minutes apart. So you get some intervals that are of length 5. You get some that are of length 10. And suppose that these are equally likely. So we have -- not exactly -- In the long run, we have as many 5 minute intervals as we have 10 minute intervals.

So the average interarrival time is 7 and 1/2. But if a person shows up at a random time, what are they going to see? Do we have as many 5s as 10s? But every 10 covers twice as much space. So if I show up at a random time, I have probability 2/3 falling inside an interval of duration 10. And I have one 1/3 probability of falling inside an interval of duration 5. That's because, out of the whole real line, 2/3 of it is covered by intervals of length 10, just because they're longer. 1/3 is covered by the smaller intervals.

Now if I fall inside an interval of length 10 and I measure the length of the interval that I fell into, that's going to be 10. But if I fall inside an interval of length 5 and I measure how long it is, I'm going to get a 5. And that, of course, is going to be different than 7.5. OK. And which number should be bigger? It's the second number that's bigger because this one is biased in favor of the longer intervals. So that's, again, another illustration of the different results that you get when you have this random incidence phenomenon.

So the bottom line, again, is that if you talk about a typical interarrival time, one must be very precise in specifying what we mean typical. So typical means sort of random. But to use the word random, you must specify very precisely what is the random experiment that you are using. And if you're not careful, you can get into apparent puzzles, such as the following. Suppose somebody tells you the average family size is 4, but the average person lives in a family of size 6. Is that compatible? Family size is 4 on the average, but typical people live, on the average, in families of size 6. Well yes. There's no contradiction here.

We're talking about two different experiments. In one experiment, I pick a family at random, and I tell you the average family is 4. In another experiment, I pick a person at random and I tell you that this person, on the average, will be in their family of size 6. And what is the catch here? That if I pick a person at random, large families are more likely to be picked. So there's a bias in favor of large families.

Or if you want to survey, let's say, are trains crowded in your city? Or are buses crowded? One choice is to pick a bus at random and inspect how crowded it is. Another choice is to pick a typical person and ask them, "Did you ride the bus today? Was it's crowded?" Well suppose that in this city there's one bus that's extremely crowded and all the other buses are completely empty. If you ask a person, "Was your bus crowded?" They will tell you, "Yes, my bus was crowded." There's no witness from the empty buses to testify in their favor. So by sampling people instead of sampling buses, you're going to get different result.

And in the process industry, if your job is to inspect and check cookies, you will be faced with a big dilemma. Do you want to find out how many chocolate chips there are on a typical cookie? Are you going to interview cookies or are you going to interview chocolate chips and ask them how many other chips where there on your cookie? And you're going to get different answers in these cases. So moral is, one has to be very precise on how you formulate the sampling procedure that you have. And you'll get different answers.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013

Transcript – Recitation: Random Incidence Under Erlang Arrivals

Hi. In this problem, we're going to look at random incidence under Erlang arrivals. First, let's parse what that means. In a Poisson process, remember, the time between arrivals, or the inter-arrival time, is distributed as an exponential random variable. And random incidence for a Poisson process refers to the somewhat surprising result that when you consider a specific time, say, T^* , then the length of the inter-arrival interval that contains that time T^* is not distributed as an exponential random variable. It's actually distributed as an Erlang random variable of order 2 or it's distributed as a sum of two exponential random variables. And the reason for that is that it comprises of two parts. One is the time since the last arrival until T^* , which is exponentially distributed, and the time from T^* until the next arrival, which is also exponentially distributed.

So that brings us to a review of what Erlang random variables are. An Erlang random variable of order k is just the sum of k independent and identically distributed exponential random variables. So to be more specific, if T_i is an exponential random variable with parameter λ , then if you take k iid copies of T_i and add them up, and call that Y_k , then Y_k is an Erlang random variable of order k .

And one other fact is that the mean of Y_k , the mean of an Erlang random variable of order k , is just k , the order, over λ , which is the rate of the underlying exponential random variables.

So as a consequence, if you have an Erlang random variable of order two and that random variable also has a mean of two over λ , we can interpret that random variable as just being the sum of two exponential random variables, 2 iid exponential random variables, T_1 and T_2 , where each one takes exponential with the rate in λ .

OK, so in this problem now, we're dealing with the random incidence not under Poisson processes, but under something else, which we call here an Erlang process with Erlang arrival times. So to be more specific, what we're saying is that, instead of inter-arrival time being exponentially distributed, in this process, and inter-arrival time is actually distributed as an Erlang random variable of order 2 with mean 2 over λ .

So to be explicit, this is no longer a Poisson process. It's some other process because the inter-arrival times are not exponential. So let's make use of this fact that we talked about earlier because now we know that the inter-arrival times of this Erlang process are Erlang order 2 with mean 2 over λ . But we know that that can just be re-interpreted as a sum of two simple exponentials, each with parameter λ .

So let's just draw another picture and imagine that for each of these arrivals, so say we have three sample arrivals in this Erlang process, we can fill in, kind of, the gaps between these with additional arrivals. And then think of each one of these times as all being exponential with parameter λ .

So this is a valid interpretation because when we connect these, these inter-arrival times correspond to the combination of two inter-arrival times, which we know we can split that into just two exponentials. So each one of these is an exponential random variable. And when you combine them, you get an Erlang order of 2.

But the nice thing about this is that if we look at this diagram, it actually is just exactly a Poisson process with a rate lambda because now, what we're dealing with are exactly-- the inter-arrival times are now exactly exponential random variables. And so this is in fact, now, just a simple Poisson process.

And we can also just think of it as we take the Poisson process, and take every other arrival, say, all the even-numbered arrivals, and make those corresponds to be arrivals in the Erlang process.

OK, so now let's think about some specific time T-star. We want to know what is the distribution of the length of this to be specific inter-arrival interval that T-star is in.

Well, what we can do is take it down to the level of this Poisson process and look at it from there. Well, we do that because, for a Poisson process, we know about random incidence for Poisson processes. And we know how to deal with Poisson processes.

So let's think about this now. Well, T-star is here. And what we know from random incidence for a Poisson processes is that the length of this inter-arrival interval for the Poisson process, we know that this is an exponential plus an exponential. So combined, this is Erlang order 2.

But that only covers from here to here. And what we want is actually from here to there. Well now, we tack on an extra exponential because we know that the inter-arrival times-- the time between this arrival and that arrival in the Poisson process is just another exponential. And now all of these are in [INAUDIBLE] time intervals. And they're all independent. And so the time of this inter-arrival interval in the Erlang process is just going to be the sum of three independent exponentials within the underlying Poisson process. And so to answer here is actually, it's going to be an Erlang of order 3.

Now this is one possible scenario for how this might occur. Another scenario is actually that T-star is somewhere else. So let's draw this again.

And suppose now, in this case, T-star landed between an even numbered arrival in the Poisson process and an odd numbered arrival. Now it could also arrive between an odd numbered and an even numbered arrival. So it could be that T-star is actually here.

Well, but in this case, it's actually more or less the same thing because now what we want is the length of this entire inter-arrival interval, which, in the Poisson world, we can break it down into random incidence within this interval, this inter-arrival interval, which is two exponentials, or an Erlang of 2, plus this interval, which is just a standard inter-arrival time, which is another exponential.

So in this case as well, we have the sum of three independent exponential random variables. And so, in either case, we have that the inter-arrival time in the Erlang process is an Erlang of order 3. And so the final answer is, in fact, that the inter-arrival for random incidence under these Erlang-type arrivals is an Erlang of order 3.

OK, so in this problem we looked at the random incidence under a different type of an arrival process, not Poisson, but with Erlang random variables. But we used the insight that Erlang really can be re-interpreted as the sum of independent and identically distributed exponential random variables. And exponential random variables can be viewed as one way of interpreting and viewing a Poisson process.

And so by going through those steps, we were able to use what we know about random incidence under Poisson processes to help us solve this problem of random incidence its Erlang arrivals. So I hope that was helpful. And I'll see you next time.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Problem Set 7
Due November 8, 2010

1. Consider a sequence of mutually independent, identically distributed, probabilistic trials. Any particular trial results in either a success (with probability p) or a failure.
 - (a) Obtain a simple expression for the probability that the i th success occurs before the j th failure. You may leave your answer in the form of a summation.
 - (b) Determine the expected value and variance of the number of successes which occur before the j th failure.
 - (c) Let L_{17} be described by a Pascal PMF of order 17. Find the numerical values of a and b in the following equation. Explain your work.
$$\sum_{l=42}^{\infty} p_{L_{17}}(l) = \sum_{x=0}^a \binom{b}{x} p^x (1-p)^{(b-x)}$$
2. Fred is giving out samples of dog food. He makes calls door to door, but he leaves a sample (one can) only on those calls for which the door is answered *and* a dog is in residence. On any call the probability of the door being answered is $3/4$, and the probability that any household has a dog is $2/3$. Assume that the events “Door answered” and “A dog lives here” are independent and also that the outcomes of all calls are independent.
 - (a) Determine the probability that Fred gives away his first sample on his third call.
 - (b) Given that he has given away exactly four samples on his first eight calls, determine the conditional probability that Fred will give away his fifth sample on his eleventh call.
 - (c) Determine the probability that he gives away his second sample on his fifth call.
 - (d) Given that he did not give away his second sample on his second call, determine the conditional probability that he will leave his second sample on his fifth call.
 - (e) We will say that Fred “needs a new supply” immediately *after* the call on which he gives away his last can. If he starts out with two cans, determine the probability that he completes at least five calls before he needs a new supply.
 - (f) If he starts out with exactly m cans, determine the expected value and variance of D_m , the number of homes with dogs which he passes up (because of no answer) before he needs a new supply.
3. Let T_1 and T_2 be exponential random variables with parameter λ , and let S be an exponential random variable with parameter μ . We assume that all three of these random variables are independent. Derive an expression for the expected value of $\min\{T_1 + T_2, S\}$. *Hint:* See Problem 6.19 in the text.
4. A single dot is placed on a very long length of yarn at the textile mill. The yarn is then cut into lengths requested by different customers. The lengths are independent of each other, but all distributed according to the PDF $f_L(\ell)$. Let R be the length of yarn purchased by that customer whose purchase included the dot. Determine the expected value of R in the following cases:

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
 (Fall 2010)

- (a) $f_L(\ell) = \lambda e^{-\lambda\ell}$, $\ell \geq 0$
- (b) $f_L(\ell) = \frac{\lambda^3 \ell^2 e^{-\lambda\ell}}{2}$, $\ell \geq 0$
- (c) $f_L(\ell) = \ell e^\ell$, $0 \leq \ell \leq 1$
5. Consider a Poisson process of rate λ . Let random variable N be the number of arrivals in $(0, t]$ and M be the number of arrivals in $(0, t + s]$, where $t, s \geq 0$.
- Find the conditional PMF of M given N , $p_{M|N}(m|n)$, for $m \geq n$.
 - Find the joint PMF of N and M , $p_{N,M}(n,m)$.
 - Find the conditional PMF of N given M , $p_{N|M}(n|m)$, for $n \leq m$, using your answer to part (b).
 - Rederive your answer to part (c) without using part (b). As a hint, consider what kind of distribution the k^{th} arrival time has if we are given the event $\{M = m\}$, where $k \leq m$.
 - Find $E[NM]$.
6. The interarrival times for cars passing a checkpoint are independent random variables with PDF

$$f_T(t) = \begin{cases} 2e^{-2t}, & \text{for } t > 0 \\ 0, & \text{otherwise.} \end{cases}$$

where the interarrival times are measured in minutes. The successive experimental values of the durations of these interarrival times are recorded on small computer cards. The recording operation occupies a negligible time period following each arrival. Each card has space for three entries. As soon as a card is filled, it is replaced by the next card.

- Determine the mean and the third moment of the interarrival times.
- Given that no car has arrived in the last four minutes, determine the PMF for random variable K , the number of cars to arrive in the next six minutes.
- Determine the PDF and the expected value for the total time required to use up the first dozen computer cards.
- Consider the following two experiments:
 - Pick a card at random from a group of completed cards and note the total time, Y , the card was in service. Find $\mathbf{E}[Y]$ and $\text{var}(Y)$.
 - Come to the corner at a random time. When the card in use at the time of your arrival is completed, note the total time it was in service (the time from the start of its service to its completion). Call this time W . Determine $\mathbf{E}[W]$ and $\text{var}(W)$.

- G1[†]. Consider a Poisson process with rate λ , and let $N(G_i)$ denote the number of arrivals of the process during an interval $G_i = (t_i, t_i + c_i]$. Suppose we have n such intervals, $i = 1, 2, \dots, n$, mutually disjoint. Denote the union of these intervals by G , and their total length by $c = c_1 + c_2 + \dots + c_n$. Given $k_i \geq 0$ and with $k = k_1 + k_2 + \dots + k_n$, determine

$$\mathbf{P}\left(N(G_1) = k_1, N(G_2) = k_2, \dots, N(G_n) = k_n \mid N(G) = k\right).$$

[†]Required for 6.431; optional for 6.041

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Problem Set 7: Solutions

1. (a) The event of the i th success occurring before the j th failure is equivalent to the i th success occurring within the first $(i + j - 1)$ trials (since the i th success must occur no later than the trial right before the j th failure). This is equivalent to event that i or more successes occur in the first $(i + j - 1)$ trials (where we can have, at most, $(i + j - 1)$ successes). Let S_i be the time of the i th success, F_j be the time of the j th failure, and N_k be the number of successes in the first k trials (so N_k is a binomial random variable over k trials). So we have:

$$\mathbf{P}(S_i < F_j) = \mathbf{P}(N_{i+j-1} \geq i) = \sum_{k=i}^{i+j-1} \binom{i+j-1}{k} p^k (1-p)^{i+j-1-k}$$

- (b) Let K be the number of successes which occur before the j th failure, and L be the number of trials to get to the j th failure. L is simply a j th order Pascal, with probability of $1 - p$ (since we are now interested in the failures, not the successes.) Plugging into the formula for j th order Pascal random variable,

$$\mathbf{E}[L] = \frac{j}{1-p}, \sigma_K^2 = \frac{p}{(1-p)^2}j$$

Since $K = L - j$,

$$\mathbf{E}[K] = \frac{p}{1-p}j, \sigma_K^2 = \frac{p}{(1-p)^2}j$$

- (c) This expression is the same as saying we need at least 42 trials to get the 17th success. Therefore, it can be rephrased as having a maximum of 16 successes in the first 41 trials. Hence $b = 41$, $a = 16$.

2. A successful call occurs with probability $p = \frac{3}{4} \cdot \frac{2}{3} = \frac{1}{2}$.

- (a) Fred will give away his first sample on the third call if the first two calls are failures and the third is a success. Since the trials are independent, the probability of this sequence of events is simply

$$(1-p)(1-p)p = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8}$$

- (b) The event of interest requires failures on the ninth and tenth trials and a success on the eleventh trial. For a Bernoulli process, the outcomes of these three trials are independent of the results of any other trials and again our answer is

$$(1-p)(1-p)p = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8}$$

- (c) We desire the probability that L_2 , the time to the second arrival is equal to five trials. We know that $p_{L_2}(\ell)$ is a Pascal PMF of order 2, and we have

$$p_{L_2}(5) = \binom{5-1}{2-1} p^2 (1-p)^{5-2} = 4 \cdot \left(\frac{1}{2}\right)^5 = \frac{1}{8}$$

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
 (Fall 2010)

- (d) Here we require the conditional probability that the experimental value of L_2 is equal to 5, given that it is greater than 2.

$$\begin{aligned}\mathbf{P}(L_2 = 5 | L_2 > 2) &= \frac{p_{L_2}(5)}{P(L_2 > 2)} = \frac{p_{L_2}(5)}{1 - p_{L_2}(2)} \\ &= \frac{\binom{5-1}{2-1} p^2 (1-p)^{5-2}}{1 - \binom{2-1}{2-1} p^2 (1-p)^0} = \frac{4 \cdot \left(\frac{1}{2}\right)^5}{1 - \left(\frac{1}{2}\right)^2} = \frac{1}{6}\end{aligned}$$

- (e) The probability that Fred will complete at least five calls before he needs a new supply is equal to the probability that the experimental value of L_2 is greater than or equal to 5.

$$\begin{aligned}\mathbf{P}(L_2 \geq 5) &= 1 - P(L_2 \leq 4) = 1 - \sum_{\ell=2}^4 \binom{\ell-1}{2-1} p^2 (1-p)^{\ell-2} \\ &= 1 - \left(\frac{1}{2}\right)^2 - \binom{2}{1} \left(\frac{1}{2}\right)^3 - \binom{3}{1} \left(\frac{1}{2}\right)^4 = \frac{5}{16}\end{aligned}$$

- (f) Let discrete random variable F represent the number of failures before Fred runs out of samples on his m th successful call. Since L_m is the number of trials up to and including the m th success, we have $F = L_m - m$. Given that Fred makes L_m calls before he needs a new supply, we can regard each of the F unsuccessful calls as trials in another Bernoulli process with parameter r , where r is the probability of a success (a disappointed dog) obtained by

$$\begin{aligned}r &= \mathbf{P}(\text{dog lives there} \mid \text{Fred did not leave a sample}) \\ &= \frac{\mathbf{P}(\text{dog lives there AND door not answered})}{1 - \mathbf{P}(\text{giving away a sample})} = \frac{\frac{1}{4} \cdot \frac{2}{3}}{1 - \frac{1}{2}} = \frac{1}{3}\end{aligned}$$

We define X to be a Bernoulli random variable with parameter r . Then, the number of dogs passed up before Fred runs out, D_m , is equal to the sum of F Bernoulli random variables each with parameter $r = \frac{1}{3}$, where F is a random variable. In other words,

$$D_m = X_1 + X_2 + X_3 + \cdots + X_F.$$

Note that D_m is a sum of a random number of independent random variables. Further, F is independent of the X_i 's since the X_i 's are defined in the conditional universe where the door is not answered, in which case, whether there is a dog or not does not affect the probability of that trial being a failed trial or not. From our results in class, we can calculate its expectation and variance by

$$\begin{aligned}\mathbf{E}[D_m] &= \mathbf{E}[F]\mathbf{E}[X] \\ \text{var}(D_m) &= \mathbf{E}[F]\text{var}(X) + (\mathbf{E}[X])^2\text{var}(F),\end{aligned}$$

where we make the following substitutions.

$$\begin{aligned}\mathbf{E}[F] &= \mathbf{E}[L_m - m] = \frac{m}{p} - m = m. \\ \text{var}(F) &= \text{var}(L_m - m) = \text{var}(L_m) = \frac{m(1-p)}{p^2} = 2m. \\ \mathbf{E}[X] &= r = \frac{1}{3}. \\ \text{var}(X) &= r(1-r) = \frac{2}{9}.\end{aligned}$$

Finally, substituting these values, we have

$$\begin{aligned}\mathbf{E}[D_m] &= m \cdot \frac{1}{3} = \frac{m}{3} \\ \text{var}(D_m) &= m \cdot \frac{2}{9} + \left(\frac{1}{3}\right)^2 \cdot 2m = \frac{4m}{9}\end{aligned}$$

3. We view the random variables T_1 and T_2 as interarrival times in two independent Poisson processes both with rate λ . S as the interarrival time in a third Poisson process (independent from the first two) with rate μ . We are interested in the expected value of the time Z until either the first process has had two arrivals or the second process has had an arrival.

Given that the first arrival was from the second process, the expected wait time for that arrival would be $\frac{1}{\mu+\lambda}$. The probability of an arrival from the second process is $\frac{\mu}{\mu+\lambda}$. Given that the first arrival time was from the first process, the expected wait time would be that for first arrival, $\frac{1}{\mu+\lambda}$, plus the expected wait time for another arrival from the merged process. Similarly, the probability of an arrival from the first process is $\frac{\lambda}{\mu+\lambda}$. Thus,

$$\begin{aligned}\mathbf{E}[Z] &= \mathbf{P}(\text{Arrival from second process})\mathbf{E}[\text{wait time}|\text{Arrival from second process}] + \\ &\quad \mathbf{P}(\text{Arrival from first process})\mathbf{E}[\text{wait time}|\text{Arrival from first process}] \\ &= \frac{\mu}{\mu+\lambda} \cdot \frac{1}{\mu+\lambda} + \frac{\lambda}{\mu+\lambda} \cdot \left(\frac{1}{\mu+\lambda} + \frac{1}{\mu+\lambda}\right).\end{aligned}$$

After some simplifications, we see that

$$\mathbf{E}[Z] = \frac{1}{\mu+\lambda} + \frac{\lambda}{\mu+\lambda} \cdot \frac{1}{\mu+\lambda}$$

4. The dot location of the yarn, as related to the size of the pieces of the yarn cut for any particular customer, can be viewed in light of the random incident paradox.

- (a) Here, the length of each piece of yarn is exponentially distributed. As explained on page 298 of the text, due to the memorylessness of the exponential, the distribution of the length of the piece of yarn containing the red dot is a second order Erlang. Thus, the $\mathbf{E}[R] = 2\mathbf{E}[L] = \frac{2}{\lambda}$.
- (b) Think of exponentially-spaced marks being made on the yarn, so the length requested by the customers each involve *three* such sections of exponentially distributed lengths (since the PDF of L is third-order Erlang). The piece of yarn with the dot will have the dot in any one of these three sections, and the expected length of that section, by (a), will be $2/\lambda$, while the expected lengths of the other two sections will be $1/\lambda$. Thus, the total expected length containing the dot is $4/\lambda$.

In general, for processes, in which the interarrival intervals with distribution $F_X(x)$ are IID, the expected length of an arbitrarily chosen interval is $\frac{\mathbf{E}[X^2]}{\mathbf{E}[X]}$. We see that for the above parts, this formula is certainly valid.

- (c) Using the formula stated above, $\mathbf{E}[L] = \int_0^1 \ell^2 e^\ell d\ell = e^\ell (\ell^2 - 2\ell + 2)|_0^1 = e - 2$
 $\mathbf{E}[L^2] = \int_0^1 \ell^3 e^\ell d\ell = e^\ell (\ell^3 - 3\ell^2 + 6\ell - 6)|_0^1 = 6 - 2e$
 Hence,

$$\mathbf{E}[R] = \boxed{\frac{6-2e}{e-2}}.$$

5. (a) We know there are n arrivals in t amount of time, so we are looking for how many extra arrivals there are in s amount of time.

$$p_{M|N}(m|n) = \frac{(\lambda s)^{m-n} e^{-\lambda s}}{(m-n)!} \quad \text{for } m \geq n \geq 0$$

(b) By definition:

$$\begin{aligned} p_{N,M}(n,m) &= p_{M|N}(m|n)p_N(n) \\ &= \frac{\lambda^m s^{m-n} t^n e^{-\lambda(s+t)}}{(m-n)!n!} \quad \text{for } m \geq n \geq 0 \end{aligned}$$

(c) By definition:

$$\begin{aligned} p_{N|M}(n|m) &= \frac{p_{M,N}(m,n)}{p_M(m)} \\ &= \binom{m}{n} \frac{s^{m-n} t^n}{(s+t)^m} \quad \text{for } m \geq n \geq 0 \end{aligned}$$

- (d) We want to find: $\mathbf{P}(N = n|M = m)$. Given $M=m$, we know that the m arrivals are uniformly distributed between 0 and $t+s$. Consider each arrival a success if it occurs before time t , and a failure otherwise. Therefore given $M=m$, N is a binomial random variable with m trials and probability of success $\frac{t}{t+s}$. We have the desired probability:

$$\mathbf{P}(N = n|M = m) = \binom{m}{n} \left(\frac{t}{t+s} \right)^n \left(\frac{s}{t+s} \right)^{m-n} \quad \text{for } m \geq n \geq 0$$

(e) We can rewrite the expectation as:

$$\begin{aligned} \mathbf{E}[NM] &= \mathbf{E}[N(M - N) + N^2] \\ &= \mathbf{E}[N]\mathbf{E}[M - N] + \mathbf{E}[N^2] \\ &= (\lambda t)(\lambda s) + (\text{var}(N) + (\mathbf{E}[N])^2) \\ &= (\lambda t)(\lambda s) + \lambda t + (\lambda t)^2 \end{aligned}$$

where the second equality is obtained via the independent increment property of the poisson process.

6. The described process for cars passing the checkpoint is a Poisson process with an arrival rate of $\lambda = 2$ cars per minute.

(a) The first and third moments are, respectively,

$$\mathbf{E}[T] = \frac{1}{\lambda} = \frac{1}{2} \quad \mathbf{E}[T^3] = \int_0^\infty t^3 2e^{-2t} dt = \frac{3!}{2^3} \underbrace{\int_0^\infty \frac{2^4 t^3 e^{-2t}}{3!} dt}_{=1} = \frac{3}{4}$$

where we recognized the integrand to be a 4th-order Erlang PDF and therefore integrating it over the entire range of the random variable must sum to unity.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
 (Fall 2010)

- (b) The Poisson process is memoryless, and thus the history of events in the previous 4 minutes does not affect the future. So, the conditional PMF for K is equivalent to the unconditional PMF that describes the number of Poisson arrivals in an interval of time, which in this case is $\tau = 6$ minutes and thus $(\lambda\tau) = 12$:

$$p_K(k) = \frac{12^k e^{12}}{k!}, \quad k = 0, 1, 2, \dots$$

- (c) The first dozen computer cards are used up upon the 36th car arrival. Letting D denote this total time, $D = T_1 + T_2 + \dots + T_{36}$, where each independent T_i is exponentially distributed with parameter $\lambda = 2$, the distribution for D is therefore a 36th-order Erlang distribution with PDF and expected value of, respectively,

$$f_D(d) = \frac{2^{36} d^{35} e^{-2d}}{35!}, \quad d \geq 0 \quad \mathbf{E}[D] = 36\mathbf{E}[T] = 18$$

- (d) In both experiments, because a card completes after registering three cars, we are considering the amount of time it takes for three cars to pass the checkpoint. In the second experiment, however, note that the manner with which the particular card is selected is biased towards cards that are in service longer. That is, the time instant at which we come to the corner is more likely to fall within a longer interarrival period – one of the three interarrival times that adds up to the total time the card is in service is selected by *random incidence* (see the end of Section 6.2 in text).

- i. The service time of any particular completed card is given by $Y = T_1 + T_2 + T_3$, and thus Y is described by a 3rd-order Erlang distribution with parameter $\lambda = 2$:

$$\mathbf{E}[Y] = \frac{3}{\lambda} = \frac{3}{2} \quad \text{var}(Y) = \frac{3}{\lambda^2} = \frac{3}{4}.$$

- ii. The service time of a particular completed card with one of the three interarrival times selected by random incidence is $W = T_1 + T_2 + L$, where L is the interarrival period that contains the time instant we arrived at the corner. Following the arguments in the text, L is Erlang of order two and thus W is described by a 4th-order Erlang distribution with parameter $\lambda = 2$:

$$\mathbf{E}[W] = \frac{4}{\lambda} = 2 \quad \text{var}(W) = \frac{4}{\lambda^2} = 1.$$

G1[†]. For simplicity, introduce the notation $N_i = N(G_i)$ for $i = 1, \dots, n$ and $N_G = N(G)$. Then

$$\begin{aligned} \mathbf{P}(N_1 = k_1, \dots, N_n = k_n | N_G = k) &= \frac{\mathbf{P}(N_1 = k_1, \dots, N_n = k_n, N_G = k)}{\mathbf{P}(N_G = k)} \\ &= \frac{\mathbf{P}(N_1 = k_1) \cdots \mathbf{P}(N_n = k_n)}{\mathbf{P}(N_G = k)} \\ &= \frac{\left(\frac{(c_1\lambda)^{k_1} e^{-c_1\lambda}}{k_1!}\right) \cdots \left(\frac{(c_n\lambda)^{k_n} e^{-c_n\lambda}}{k_n!}\right)}{\left(\frac{(c\lambda)^k e^{-c\lambda}}{k!}\right)} \\ &= \frac{k!}{k_1! \cdots k_n!} \left(\frac{c_1}{c}\right)^{k_1} \cdots \left(\frac{c_n}{c}\right)^{k_n} \\ &= \binom{k}{k_1 \ \dots \ k_n} \left(\frac{c_1}{c}\right)^{k_1} \cdots \left(\frac{c_n}{c}\right)^{k_n} \end{aligned}$$

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Fall 2010)

The result can be interpreted as a *multinomial distribution*. Imagine we throw an n-sided die k times, where Side i comes up with probability $p_i = c_i/c$. The probability that side i comes up k_i times is given by the expression above. Now relating it back to the Poisson process that we have, each side corresponds to an interval that we sample, and the probability that we sample it depends directly on its relative length. This is consistent with the intuition that, given a number of Poisson arrivals in a specified interval, the arrivals are uniformly distributed.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

LECTURE 15

Poisson process — II

- **Readings:** Finish Section 6.2.

- Review of Poisson process
- Merging and splitting
- Examples
- Random incidence

Review

- Defining characteristics
 - **Time homogeneity:** $P(k, \tau)$
 - **Independence**
 - **Small interval probabilities** (small δ):

$$P(k, \delta) \approx \begin{cases} 1 - \lambda\delta, & \text{if } k = 0, \\ \lambda\delta, & \text{if } k = 1, \\ 0, & \text{if } k > 1. \end{cases}$$

- N_τ is a Poisson r.v., with parameter $\lambda\tau$:

$$P(k, \tau) = \frac{(\lambda\tau)^k e^{-\lambda\tau}}{k!}, \quad k = 0, 1, \dots$$

$$\mathbf{E}[N_\tau] = \text{var}(N_\tau) = \lambda\tau$$

- Interarrival times ($k = 1$): exponential:

$$f_{T_1}(t) = \lambda e^{-\lambda t}, \quad t \geq 0, \quad \mathbf{E}[T_1] = 1/\lambda$$

- Time Y_k to k th arrival: Erlang(k):

$$f_{Y_k}(y) = \frac{\lambda^k y^{k-1} e^{-\lambda y}}{(k-1)!}, \quad y \geq 0$$

Poisson fishing

- Assume: Poisson, $\lambda = 0.6/\text{hour}$.
 - Fish for two hours.
 - if no catch, continue until first catch.

a) $\mathbf{P}(\text{fish for more than two hours}) =$

b) $\mathbf{P}(\text{fish for more than two and less than five hours}) =$

c) $\mathbf{P}(\text{catch at least two fish}) =$

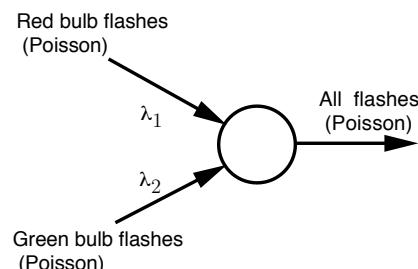
d) $\mathbf{E}[\text{number of fish}] =$

e) $\mathbf{E}[\text{future fishing time} \mid \text{fished for four hours}] =$

f) $\mathbf{E}[\text{total fishing time}] =$

Merging Poisson Processes (again)

- Merging of independent Poisson processes is Poisson



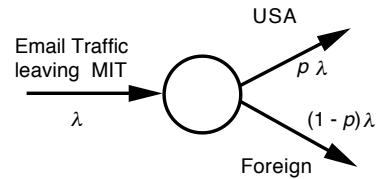
- What is the probability that the next arrival comes from the first process?

Light bulb example

- Each light bulb has independent, $\text{exponential}(\lambda)$ lifetime
- Install three light bulbs.
Find expected time until last light bulb dies out.

Splitting of Poisson processes

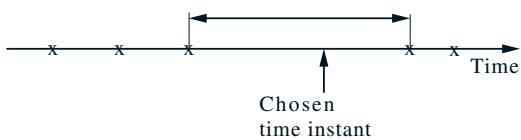
- Assume that email traffic through a server is a Poisson process.
Destinations of different messages are independent.



- Each output stream is Poisson.

Random incidence for Poisson

- Poisson process that has been running forever
- Show up at some “random time”
(really means “arbitrary time”)



- What is the distribution of the length of the chosen interarrival interval?

Random incidence in “renewal processes”

- Series of successive arrivals
 - i.i.d. interarrival times
(but not necessarily exponential)
- **Example:**
Bus interarrival times are equally likely to be 5 or 10 minutes
- If you arrive at a “random time”:
 - what is the probability that you selected a 5 minute interarrival interval?
 - what is the expected time to next arrival?

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Recitation 17
November 4, 2010

1. Iwana Passe is taking a multiple-choice exam. You may assume that the number of questions is infinite. *Simultaneously, but independently*, her conscious and subconscious faculties are generating answers for her, each in a Poisson manner. (Her conscious and subconscious are always working on different questions.) Conscious responses are generated at the rate λ_c responses per minute. Subconscious responses are generated at the rate λ_s responses per minute. Assume $\lambda_c \neq \lambda_s$. Each conscious response is an independent Bernoulli trial with probability p_c of being correct. Similarly, each subconscious response is an independent Bernoulli trial with probability p_s of being correct. Iwana responds only once to each question, and you can assume that her time for recording these conscious and subconscious responses is negligible.
 - (a) Determine $p_K(k)$, the probability mass function for the number of *conscious responses* Iwana makes in an interval of T minutes.
 - (b) If we pick any question to which Iwana has responded, what is the probability that her answer to that question:
 - i. Represents a conscious response
 - ii. Represents a conscious correct response
 - (c) If we pick an interval of T minutes, what is the probability that in that interval Iwana will make exactly r conscious responses *and* s subconscious responses?
 - (d) Determine the probability density function for random variable X , where X is the time from the start of the exam until Iwana makes her first conscious response which is preceded by at least one subconscious response.
2. Shem, a local policeman, drives from intersection to intersection in times that are independent and all exponentially distributed with parameter λ . At each intersection he observes (and reports) a car accident with probability p . (This activity does not slow his driving at all.) Independently of all else, Shem receives extremely brief radio calls in a Poisson manner with an average rate of μ calls per hour.
 - (a) Determine the PMF for N , the number of intersections Shem visits up to and including the one where he reports his first accident.
 - (b) Determine the PDF for Q , the length of time Shem drives between reporting accidents.
 - (c) What is the PMF for M , the number of accidents which Shem reports in two hours?
 - (d) What is the PMF for K , the number of accidents Shem reports between his receipt of two successive radio calls?
 - (e) We observe Shem at a random instant long after his shift has begun. Let W be the total time from Shem's last radio call until his next radio call. What is the PDF of W ?
3. Problem 6.27, page 337 in the textbook. **Random incidence in an Erlang arrival process.** Consider an arrival process in which the interarrival times are independent Erlang random variables of order 2, with mean $2/\lambda$. Assume that the arrival process has been ongoing for a very long time. An external observer arrives at a given time t . Find the PDF of the length of the interarrival interval that contains t .

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Recitation 17 Solutions

November 4, 2010

1. (a) K has a Poisson distribution with average arrival time $\mu = \lambda_c T$

$$p_K(k) = \frac{(\lambda_c T)^k e^{-\lambda_c T}}{k!}, \quad k = 0, 1, 2, \dots; T \geq 0.$$

- (b) i. $\mathbf{P}(\text{conscious response}) = \left(\frac{\lambda_c}{\lambda_c + \lambda_s} \right).$
 ii. $\mathbf{P}(\text{conscious correct response}) = \mathbf{P}(\text{conscious resp}) \mathbf{P}(\text{correct resp} | \text{conscious resp}) = \left(\frac{\lambda_c}{\lambda_c + \lambda_s} p_c \right).$
 (c) Since the conscious and subconscious responses are generated independently,

$$\begin{aligned} & \mathbf{P}(r \text{ conscious responses and } s \text{ subconscious responses in interval } T) \\ &= \mathbf{P}(r \text{ conscious responses in } T) \mathbf{P}(s \text{ subconscious responses in } T) \\ &= \frac{(\lambda_c T)^r e^{-\lambda_c T}}{r!} \cdot \frac{(\lambda_s T)^s e^{-\lambda_s T}}{s!} \end{aligned}$$

- (d) Let X_s = the time from the start of the exam to the time of the 1st subconscious response, and X_c = the time from the 1st subconscious response to the time of the next conscious response.

Note that X_s and X_c are independent exponentially distributed random variables with parameters λ_s and λ_c , respectively.

$$\begin{aligned} f_{X_s}(x_s) &= \lambda_s e^{-\lambda_s x_s} \text{ when } x_s \geq 0 \\ &= 0 \text{ otherwise} \\ f_{X_c}(x_c) &= \lambda_c e^{-\lambda_c x_c} \text{ when } x_c \geq 0 \\ &= 0 \text{ otherwise} \end{aligned}$$

$X = X_s + X_c$. So its PDF is the convolution of the two exponential distributions. For $x \geq 0$

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} \lambda_s e^{-\lambda_s(x-x_c)} \lambda_c e^{-\lambda_c x_c} dx_c \\ &= \int_0^x \lambda_s \lambda_c e^{-\lambda_s x} e^{(\lambda_s - \lambda_c)x_c} dx_c \text{ because } x - x_c > 0 \\ &= \lambda_s \lambda_c e^{-\lambda_s x} \int_0^x e^{(\lambda_s - \lambda_c)x_c} dx_c \\ &= \frac{\lambda_s \lambda_c}{\lambda_s - \lambda_c} e^{-\lambda_s x} (e^{(\lambda_s - \lambda_c)x} - 1) \text{ because } \lambda_s \neq \lambda_c \\ &= \frac{\lambda_s \lambda_c}{\lambda_s - \lambda_c} (e^{-\lambda_c x} - e^{-\lambda_s x}) \end{aligned}$$

2. (a) Since we are looking for the number of “trials” up to and including the first “success,” N is a geometric random variable with parameter p .

$$p_N(n) = (1-p)^{n-1} p, \quad n \geq 1.$$

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Fall 2010)

- (b) The length of time spent driving to each intersection is exponentially distributed with parameter λ . Since the probability of Shem observing an accident at a given intersection is p , the distribution of the length of time in between accident reports is exponential but with parameter $p\lambda$ (think of Poisson splitting). Thus,
$$f_Q(q) = (p\lambda)e^{-qp\lambda}, q \geq 0.$$
- (c) Since the interarrival time of accidents is exponentially distributed with parameter $p\lambda$, the number of arrivals in a given amount of time τ is a Poisson random variable with parameter $p\lambda\tau$. Thus,
$$\mathbf{P}(m \text{ arrivals in } 2 \text{ hours}) = p_M(m) = \frac{e^{-2p\lambda}(2p\lambda)^m}{m!}, \quad m \geq 0.$$
- (d) We can view the radio calls to Shem and the accident reports as independent Poisson processes with arrival rates μ and $p\lambda$, respectively. When the two independent Poisson processes are joined, the resultant is a Poisson process with arrival rate $\mu+p\lambda$. Furthermore, the probability of an arrival from the radio calls is $\frac{\mu}{\mu+p\lambda}$. Since we are interested in the number of reported accidents between two radio calls, we can view this as a shifted Geometric random variable with parameter $\frac{\mu}{\mu+p\lambda}$. Thus,
$$p_K(k) = \left(\frac{p\lambda}{\mu+p\lambda}\right)^k \left(\frac{\mu}{\mu+p\lambda}\right), \quad k \geq 0.$$
- (e) If we begin to observe Shem's radio calls at some random instant in time, due to the memoryless property of Poisson interarrivals, the distribution until he receives the next call will still be exponential with parameter μ . Also, the time from the previous call until the point at which we begin to observe Shem is also an exponential distribution with parameter μ . Thus, $W = X_1 + X_2$, where X_1 and X_2 have exponential distributions, i.e. W is a second order Erlang PDF.
$$f_W(w) = (\mu)^2 w e^{-w\mu}$$
3. See problem 6.27, page 337 in the textbook.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Tutorial 8
November 4/5, 2010

1. Type A, B, and C items are placed in a common buffer, each type arriving as part of an independent Poisson process with average arrival rates, respectively, of a , b , and c items per minute. For the first four parts of this problem, assume the buffer is discharged immediately whenever it contains a total of ten items.
 - (a) What is the probability that, of the first ten items to arrive at the buffer, only the first and one other are type A?
 - (b) What is the probability that any particular discharge of the buffer contains five times as many type A items as type B items?
 - (c) Determine the PDF, expectation, and variance for the total time between consecutive discharges of the buffer.
 - (d) Determine the probability that exactly two of each of the three item types arrive at the buffer input during any particular five minute interval.
2. A store opens at $t = 0$ and *potential* customers arrive in a Poisson manner at an average arrival rate of λ potential customers per hour. As long as the store is open, and independently of all other events, each particular potential customer becomes an *actual* customer with probability p . The store closes as soon as ten actual customers have arrived.
 - (a) What is the probability that exactly three of the first five potential customers become actual customers?
 - (b) What is the probability that the fifth potential customer to arrive becomes the third actual customer?
 - (c) What is the PDF and expected value for L , the duration of the interval from store opening to store closing?
 - (d) Given only that exactly three of the first five potential customers became actual customers, what is the conditional expected value of the *total* time the store is open?
 - (e) Considering only customers arriving between $t = 0$ and the closing of the store, what is the probability that no two *actual* customers arrive within τ time units of each other?
3. Problem 6.24, page 335 in text.

Consider a Poisson process with parameter λ , and an independent random variable T , which is exponential with parameter ν . Find the PMF of the number of Poisson arrivals during the time interval $[0, T]$.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Tutorial 8: Solutions

1. (a) Let $A = \{\text{An arriving item is of type A}\}$.

The 3 independent Poisson processes can be merged into one Poisson process with an arrival rate of $a + b + c$ items per minute. If δ is a short time interval, then

$$\begin{aligned} \mathbf{P}(A) &= \mathbf{P}(\{\text{A type A item arrives}\} | \{\text{An item arrives}\}) \\ &= \frac{\mathbf{P}(\{\text{A type A item arrives}\})}{\mathbf{P}(\{\text{An item arrives}\})} \\ &\approx \frac{a\delta}{(a+b+c)\delta} \\ &= \frac{a}{a+b+c}. \end{aligned}$$

The probability that the first item is type A and exactly one of the next 9 items is type A is

$$\mathbf{P}(A) \binom{9}{1} \mathbf{P}(A)(1 - \mathbf{P}(A))^8 = \binom{9}{1} \mathbf{P}(A)^2 (1 - \mathbf{P}(A))^8.$$

- (b) Let $B = \{\text{An arriving item is of type B}\}$.

Let $C = \{\text{An arriving item is of type C}\}$.

In order for there to be 5 times as many type A items as type B items, there can be either 5 type A items, 1 type B item, and 4 type C items, or 0 type A items, 0 type B item and 10 type C items. The probability of this event is the sum of the probabilities of the two cases:

$$\frac{10!}{5!1!4!} \mathbf{P}(A)^5 \mathbf{P}(B) \mathbf{P}(C)^4 + \mathbf{P}(C)^{10}$$

where $\mathbf{P}(A) = \frac{a}{a+b+c}$, $\mathbf{P}(B) = \frac{b}{a+b+c}$, and $\mathbf{P}(C) = \frac{c}{a+b+c}$.

- (c) The total time between consecutive discharges is an Erlang random variable of order 10 and parameter $a + b + c$, with PDF

$$f_T(t) = \begin{cases} \frac{(a+b+c)^{10} t^9 e^{-(a+b+c)t}}{9!} & t \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

The mean and variance are $\mathbf{E}[T] = \frac{10}{a+b+c}$ and $\text{var}(T) = \frac{10}{(a+b+c)^2}$ respectively.

- (d) Let $X = \{\text{Exactly two items from A arrive in 5 minutes}\}$.

Let $Y = \{\text{Exactly two items from B arrive in 5 minutes}\}$.

Let $Z = \{\text{Exactly two items from C arrive in 5 minutes}\}$.

Because X, Y, Z are independent events, we know that

$$\begin{aligned} \mathbf{P}(\{\text{Exactly two of each of the three types arrive in 5 minutes}\}) &= \mathbf{P}(X)\mathbf{P}(Y)\mathbf{P}(Z) \\ &= \left(\frac{(5a)^2 e^{-5a}}{2}\right) \left(\frac{(5b)^2 e^{-5b}}{2}\right) \left(\frac{(5c)^2 e^{-5c}}{2}\right) \\ &= \frac{(125abc)^2 e^{-5(a+b+c)}}{8}. \end{aligned}$$

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Fall 2010)

2. (a) Since a given potential customer becomes an actual customer with probability p , we are dealing with a binomial probability, and $\binom{5}{3}p^3(1-p)^2$ is the probability of exactly 3 of the first 5 potential customers being actual customers.
- (b) Potential customers become actual customers independently of each other. Thus,
$$\begin{aligned} & \mathbf{P}(\text{the fifth customer is the third actual customer}) \\ &= \mathbf{P}(\text{any 2 of the first 4 customers become actual customer}) \\ &\cdot \mathbf{P}(\text{the fifth customer becomes an actual customer}) \\ &= [\binom{4}{2}p^2(1-p)^2][p] = \binom{4}{2}p^3(1-p)^2. \end{aligned}$$
- (c) The process of incoming customers is being randomly split into two independent Poisson processes: the process for the arrival of actual customers, with rate $p\lambda$, and the process for the arrival of potential customers who do *not* become actual customers, with rate $(1-p)\lambda$. The store will close when 10 actual customers have arrived. Thus, the store closes when we have 10 arrivals from the Poisson process of actual customers, which has rate $p\lambda$. Thus, $L = T_1 + T_2 + \dots + T_{10}$, where T_i is the interarrival time with an exponential distribution and parameter $p\lambda$. The distribution of L is the 10th order Erlang PDF,
$$f_L(l) = \frac{(p\lambda)^{10}l^9e^{-(p\lambda)l}}{9!}, l \geq 0.$$
 It follows that expected value of L is $\frac{10}{p\lambda}$.
- (d) Since five potential customers have arrived, three of which are actual customers, we are interested in the time for the next seven actual customers to arrive. Following from part (c), the expected time for the next seven actual customers to arrive is $\frac{7}{p\lambda}$. Adding the expected time for the first five potential customers to arrive, we get that the conditional expectation for the total time the store is open is $\frac{5}{\lambda} + \frac{7}{p\lambda}$.
- (e) The probability of no two actual customers arriving within τ time units of each other is equivalent to the probability of all nine independent interarrival times, separating the ten actual customers, being at least τ time units apart. Thus,
$$\begin{aligned} & \mathbf{P}(\text{no two actual customers arriving within } \tau \text{ time units of each other}) = \\ & \mathbf{P}([T_1 \geq \tau]^9) = e^{-9p\lambda\tau}. \end{aligned}$$
3. See problem 6.24, page 334 in the textbook.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013

Transcript – Lecture 16

The following content is provided under a Creative Commons license. Your support will help MIT OpenCourseWare continue to offer high quality educational resources for free. To make a donation or view additional materials from hundreds of MIT courses, visit MIT OpenCourseWare at ocw.mit.edu.

PROFESSOR: So we're going to start now with a new chapter. We're going to talk about Markov processes. The good news is that this is a subject that is a lot more intuitive and simple in many ways than, let's say, the Poisson processes. So hopefully this will be enjoyable.

So Markov processes is, a general class of random processes. In some sense, it's more elaborate than the Bernoulli and Poisson processes, because now we're going to have dependencies between difference times, instead of having memoryless processes. So the basic idea is the following. In physics, for example, you write down equations for how a system evolves that has the general form. The new state of a system one second later is some function of old state.

So Newton's equations and all that in physics allow you to write equations of this kind. And so if that a particle is moving at a certain velocity and it's at some location, you can predict when it's going to be a little later. Markov processes have the same flavor, except that there's also some randomness thrown inside the equation. So that's what Markov process essentially is. It describes the evolution of the system, or some variables, but in the presence of some noise so that the motion itself is a bit random.

So this is a pretty general framework. So pretty much any useful or interesting random process that you can think about, you can always described it as a Markov process if you define properly the notion of the state. So what we're going to do is we're going to introduce the class of Markov processes by, example, by talking about the checkout counter in a supermarket. Then we're going to abstract from our example so that we get a more general definition. And then we're going to do a few things, such as how to predict what's going to happen n time steps later, if we start at the particular state. And then talk a little bit about some structural properties of Markov processes or Markov chains.

So here's our example. You go to the checkout counter at the supermarket, and you stand there and watch the customers who come. So customers come, they get in queue, and customers get served one at a time. So the discussion is going to be in terms of supermarket checkout counters, but the same story applies to any service system. You may have a server, jobs arrive to that server, they get put into the queue, and the server processes those jobs one at a time.

Now to make a probabilistic model, we need to make some assumption about the customer arrivals and the customer departures. And we want to keep things as simple as possible to get started. So let's assume that customers arrive according to a Bernoulli process with some

parameter b . So essentially, that's the same as the assumption that the time between consecutive customer arrivals is a geometric random variable with parameter b .

Another way of thinking about the arrival process-- that's not how it happens, but it's helpful, mathematically, is to think of someone who's flipping a coin with bias equal to b . And whenever the coin lands heads, then a customer arrives. So it's as if there's a coin flip being done by nature that decides the arrivals of the customers. So we know that coin flipping to determine the customer arrivals is the same as having geometric inter-arrival times. We know that from our study of the Bernoulli process.

OK. And now how about the customer service times. We're going to assume that-- OK. If there is no customer in queue, no one being served, then of course, no one is going to depart from the queue. But if there a customer in queue, then that customer starts being served, and is going to be served for a random amount of time. And we make the assumption that the time it takes for the clerk to serve the customer has a geometric distribution with some known parameter q .

So the time it takes to serve a customer is random, because it's random how many items they got in their cart, and how many coupons they have to unload and so on. So it's random. In the real world, it has some probability distribution. Let's not care exactly about what it would be in the real world, but as a modeling approximation or just to get started, let's pretend that customer service time are well described by a geometric distribution, with a parameter q .

An equivalent way of thinking about the customer service, mathematically, would be, again, in terms of coin flipping. That is, the clerk has a coin with a bias, and at each time slot the clerk flips the coin. With probability q , service is over. With probability $1-q$, you continue the service process.

An assumption that we're going to make is that the coin flips that happen here to determine the arrivals, they're all independent of each other. The coin flips that determine the end of service are also independent from each other. But also the coin flips involved here are independent from the coin flips that happened there. So how arrivals happen is independent with what happens at the service process.

OK. So suppose now you want to answer a question such as the following. The time is 7:00 PM. What's the probability that the customer will be departing at this particular time? Well, you say, it depends. If the queue is empty at that time, then you're certain that you're not going to have a customer departure. But if the queue is not empty, then there is probability q that a departure will happen at that time.

So the answer to a question like this has something to do with the state of the system at that time. It depends what the queue is. And if I ask you, will the queue be empty at 7:10? Well, the answer to that question depends on whether at 7 o'clock whether the queue was huge or not. So knowing something about the state of the queue right now gives me relevant information about what may happen in the future.

So what is the state of the system? Therefore we're brought to start using this term. So the state basically corresponds to anything that's relevant. Anything that's happening right now that's kind of relevant to what may happen in the future. Knowing the size of the queue right now, is useful information for me to make predictions about what may happen 2 minutes later from now. So in this particular example, a reasonable choice for the state is to just count how many customers we have in the queue. And let's assume that our supermarket building is not too big, so it can only hold 10 people.

So we're going to limit the states. Instead of going from 0 to infinity, we're going to truncate our model at ten. So we have 11 possible states, corresponding to 0 customers in queue, 1 customer in queue, 2 customers, and so on, all the way up to 10. So these are the different possible states of the system, assuming that the store cannot handle more than 10 customers. So this is the first step, to write down the set of possible states for our system. Then the next thing to do is to start describing the possible transitions between the states.

At any given time step, what are the things that can happen? We can have a customer arrival, which moves the state 1 higher. We can have a customer departure, which moves the state 1 lower. There's a possibility that nothing happens, in which case the state stays the same. And there's also the possibility of having simultaneously an arrival and a departure, in which case the state again stays the same.

So let's write some representative probabilities. If we have 2 customers, the probability that during this step we go down, this is the probability that we have a service completion, but to no customer arrival. So this is the probability associated with this transition. The other possibility is that there's a customer arrival, which happens with probability p , and we do not have a customer departure, and so the probability of that particular transition is this number.

And then finally, the probability that we stay in the same state, this can happen in 2 possible ways. One way is that we have an arrival and a departure simultaneously. And the other possibility is that we have no arrival and no departure, so that the state stays the same. So these transition probabilities would be the same starting from any other states, state 3, or state 9, and so on. Transition probabilities become a little different at the borders, at the boundaries of this diagram, because if you're in a state 0, then you cannot have any customer departures.

There's no one to be served, but there is a probability p that the customer arrives, in which case the number of customers in the system goes to 1. Then probability $1-p$, nothing happens. Similarly with departures, if the system is full, there's no room for another arrival. But we may have a departure that happens with probability q , and nothing happens with probability $1-q$. So this is the full transition diagram annotated with transition probabilities.

And this is a complete description of a discrete time, finite state Markov chain. So this is a complete probabilistic model. Once you have all of these pieces of information, you can start calculating things, and trying to predict what's going to happen in the future. Now let us abstract from this example and come up with a more general definition. So we have this concept of the state which describes the current situation in the system that we're looking at.

The current state is random, so we're going to think of it as a random variable X_n is the state, and transitions after the system started operating. So the system starts operating at some initial state X_0 , and after n transitions, it moves to state X_n . Now we have a set of possible states. State 1 state 2, state 3, and in general, state i and state j . To keep things simple, we assume that the set of possible states is a finite set.

As you can imagine, we can have systems in which the state space is going to be infinite. It could be discrete, or continuous. But all that is more difficult and more complicated. It makes sense to start from the simplest possible setting where we just deal with the finite state space.

And time is discrete, so we can think of this state in the beginning, after 1 transition, 2 transitions, and so on. So we're in discrete time and we have finite in many states. So the system starts somewhere, and at every time step, the state is, let's say, here. A whistle blows, and the state jumps to a random next state. So it may move here, or it may move there, or it may move here, or it might stay in the place. So one possible transition is the transition before you jump, and just land in the same place where you started from.

Now we want to describe the statistics of these transitions. If I am at that state, how likely is it to that, next time, I'm going to find myself at that state? Well, we describe the statistics of this transition by writing down a transition probability, the transition probability of going from state 3 to state 1. So this transition probability is to be thought of as a conditional probability. Given that right now I am at state i what is the probability that next time I find myself at state j ? So given that right now I am at state 3, P_{31} is the probability that the next time I'm going to find myself at state 1.

Similarly here, we would have a probability P_{3i} , which is the probability that given that right now I'm at state 3, next time I'm going to find myself at state i . Now one can write such conditional probabilities down in principle, but we need to make-- so you might think of this as a definition here, but we need to make one additional big assumption, and this is the assumption that to make a process to be a Markov process. This is the so-called Markov property, and here's what it says. Let me describe it first in words here.

Every time that I find myself at state 3, the probability that next time I'm going to find myself at state 1 is this particular number, no matter how I got there. That is, this transition probability is not affected by the past of the process. It doesn't care about what path I used to find myself at state 3.

Mathematically, it means the following. You have this transition probability that from state i jump to state j . Suppose that I gave you some additional information, that I told you everything else that happened in the past of the process, everything that happened, how did you get to state i ? The assumption we're making is that this information about the past has no bearing in making predictions about the future, as long as you know where you are right now. So if I tell you, right now, you are at state i , and by the way, you got there by following a particular path, you can ignore the extra information of the particular path that you followed.

You only take into account where you are right now. So every time you find yourself at that state, no matter how you got there, you will find yourself next time at state 1 with probability P₃₁. So the past has no bearing into the future, as long as you know where you are sitting right now. For this property to happen, you need to choose your state carefully in the right way. In that sense, the states needs to include any information that's relevant about the future of the system. Anything that's not in the state is not going to play a role, but the state needs to have all the information that's relevant in determining what kind of transitions are going to happen next.

So to take an example, before you go to Markov process, just from the deterministic world, if you have a ball that's flying up in the air, and you want to make predictions about the future. If I tell you that the state of the ball is the position of the ball at the particular time, is that enough for you to make predictions where the ball is going to go next? No. You need to know both the position and the velocity. If you know position and velocity, you can make predictions about the future.

So the state of a ball that's flying is position together with velocity. If you were to just take position, that would not be enough information, because if I tell you current position, and then I tell you past position, you could use the information from the past position to complete the trajectory and to make the prediction. So information from the past is useful if you don't know the velocity. But if both position and velocity, you don't care how you got there, or what time you started. From position and velocity, you can make predictions about the future.

So there's a certain art, or a certain element of thinking, a non-mechanical aspect into problems of this kind, to figure out which is the right state variable. When you define the state of your system, you need to define it in such a way that includes all information that has been accumulated that has some relevance for the future. So the general process for coming up with a Markov model is to first make this big decision of what your state variable is going to be. Then you write down if it may be a picture of the different states. Then you identify the possible transitions.

So sometimes the diagram that you're going to have will not include all the possible arcs. You would only show those arcs that correspond to transitions that are possible. For example, in the supermarket example, we did not have a transition from state 2 to state 5, because that cannot happen. You can only have 1 arrival at any time. So in the diagram, we only showed the possible transitions.

And for each of the possible transitions, then you work with the description of the model to figure out the correct transition probability. So you got the diagram by writing down transition probabilities.

OK, so suppose you got your Markov model. What will you do with it? Well, what do we need models for? We need models in order to make predictions, to make probabilistic predictions. So for example, I tell you that the process started in that state. You let it run for some time. Where do you think it's going to be 10 time steps from now? That's a question that you might want to answer.

Since the process is random, there's no way for you to tell me exactly where it's going to be. But maybe you can give me probabilities. You can tell me, with so much probability, the state would be there. With so much probability, the state would be there, and so on. So our first exercise is to calculate those probabilities about what may happen to the process a number of steps in the future. It's handy to have some notation in here.

So somebody tells us that this process starts at the particular state i . We let the process run for n transitions. It may land at some state j , but that state j at which it's going to land is going to be random. So we want to give probabilities. Tell me, with what probability the state, n times steps later, is going to be that particular state j ?

The shorthand notation is to use this symbol here for the n -step transition probabilities that you find yourself at state j given that you started at state i . So the way these two indices are ordered, the way to think about them is that from i , you go to j . So the probability that from i you go to j if you have n steps in front of you. Some of these transition probabilities are, of course easy to write. For example, in 0 transitions, you're going to be exactly where you started. So this probability is going to be equal to 1 if i is equal to j , And 0 if i is different than j .

That's an easy one to write down. If you have only 1 transition, what's the probability that 1 step later you find yourself in state j given that you started at state i ? What is this? These are just the ordinary 1-step transition probabilities that we are given in the description of the problem. So by definition, the 1-step transition probabilities are of this form. This equality is correct just because of the way that we defined those two quantities.

Now we want to say something about the n -step transition probabilities when n is a bigger number. OK. So here, we're going to use the total probability theorem. So we're going to condition in two different scenarios, and break up the calculation of this quantity, by considering the different ways that this event can happen. So what is the event of interest? The event of interest is the following. At time 0 we start i . We are interested in landing at time n at the particular state j .

Now this event can happen in several different ways, in lots of different ways. But let us group them into subgroups. One group, or one sort of scenario, is the following. During the first $n-1$ time steps, things happen, and somehow you end up at state 1. And then from state 1, in the next time step you make a transition to state j . This particular arc here actually corresponds to lots and lots of different possible scenarios, or different spots, or different transitions. In $n-1$ time steps, there's lots of possible ways by which you could end up at state 1. Different paths through the state space.

But all of them together collectively have a probability, which is the $(n-1)$ -step transition probability, that from state i , you end up at state 1

And then there's other possible scenarios. Perhaps in the first $n-1$ time steps, you follow the trajectory that took you at state m . And then from state m , you did this transition, and you ended up at state j . So this diagram breaks up the set of all possible trajectories from i to j into different

collections, where each collection has to do with which one happens to be the state just before the last time step, just before time n . And we're going to condition on the state at time $n-1$.

So the total probability of ending up at state j is the sum of the probabilities of the different scenarios -- the different ways that you can get to state j . If we look at that type of scenario, what's the probability of that scenario happening? With probability $R_{i1}(n-1)$, I find myself at state 1 at time $n-1$. This is just by the definition of these multi-step transition probabilities. This is the number of transitions.

The probability that from state i , I end up at state 1. And then given that I found myself at state 1, with probability P_{1j} , that's the transition probability, next time I'm going to find myself at state j . So the product of these two is the total probability of my getting from state i to state j through state 1 at the time before. Now where exactly did we use the Markov assumption here? No matter which particular path we used to get from i to state 1, the probability that next I'm going to make this transition is that same number, P_{1j} .

So that number does not depend on the particular path that I followed in order to get there. If we didn't have the Markov assumption, we should have considered all possible individual trajectories here, and then we would need to use the transition probability that corresponds to that particular trajectory. But because of the Markov assumption, the only thing that matters is that right now we are at state 1. It does not matter how we got there.

So now once you see this scenario, then this scenario, and that scenario, and you add the probabilities of these different scenarios, you end up with this formula here, which is a recursion. It tells us that once you have computed the $(n-1)$ -step transition probabilities, then you can compute also the n -step transition probabilities. This is a recursion that you execute or you run for all i 's and j 's simultaneously. That is fixed. And for a particular n , you calculate this quantity for all possible i 's, j 's, k 's. You have all of those quantities, and then you use this equation to find those numbers again for all the possible i 's and j 's.

Now this is formula which is always true, and there's a big idea behind the formula. And now there's variations of this formula, depending on whether you're interested in something that's slightly different. So for example, if you were to have a random initial state, somebody gives you the probability distribution of the initial state, so you're told that with probability such and such, you're going to start at state 1. With that probability, you're going to start at state 2, and so on. And you want to find the probability at the time n you find yourself at state j .

Well again, total probability theorem, you condition on the initial state. With this probability you find yourself at that particular initial state, and given that this is your initial state, this is the probability that n time steps later you find yourself at state j . Now building again on the same idea, you can run every recursion of this kind by conditioning at different times. So here's a variation. You start at state i . After 1 time step, you find yourself at state 1, with probability p_{i1} , and you find yourself at state m with probability P_{im} . And once that happens, then you're going to follow some trajectories. And there is a possibility that you're going to end up at state j after $n-1$ time steps.

This scenario can happen in many possible ways. There's lots of possible paths from state 1 to state j. There's many paths from state 1 to state j. What is the collective probability of all these transitions? This is the event that, starting from state 1, I end up at state j in n-1 time steps. So this one has here probability R_{1j} of $n-1$. And similarly down here. And then by using the same way of thinking as before, we get the formula that $R_{ij}(n)$ is the sum over all k's of P_{ik} , and then the $R_{kj}(n-1)$.

So this formula looks almost the same as this one, but it's actually different. The indices and the way things work out are a bit different, but the basic idea is the same. Here we use the total probability theory by conditioning on the state just 1 step before the end of our time horizon. Here we use total probability theorem by conditioning on the state right after the first transition. So this generally idea has different variations. They're all valid, and depending on the context that you're dealing with, you might want to work with one of these or another.

So let's illustrate these calculations in terms of an example. So in this example, we just have 2 states, and somebody gives us transition probabilities to be those particular numbers. Let's write down the equations. So the probability that starting from state 1, I find myself at state 1 n time steps later. This can happen in 2 ways. At time $n-1$, I might find myself at state 2. And then from state 2, I make a transition back to state 1, which happens with probability-- why'd I put 2 there -- anyway, 0.2. And another way is that from state 1, I go to state 1 in $n-1$ steps, and then from state 1 I stay where I am, which happens with probability 0.5.

So this is for $R_{11}(n)$. Now $R_{12}(n)$, we can write a similar recursion for this one. On the other hand, seems these are probabilities. The state at time n is going to be either state 1 or state 2. So these 2 numbers need to add to 1, so we can just write this as $1 - R_{11}(n)$. And this is an enough of a recursion to propagate R_{11} and R_{12} as time goes on. So after $n-1$ transitions, either I find myself in state 2, and then there's a point to transition that I go to 1, or I find myself in state 1, which with that probability, and from there, I have probability 0.5 of staying where I am.

Now let's start calculating. As we discussed before, if I start at state 1, after 0 transitions I'm certain to be at state , and I'm certain not to be at state 1. If I start from state 1, I'm certain to not to be at state at that time, and I'm certain that I am right now, it's state 1. After I make transition, starting from state 1, there's probability 0.5 that I stay at state 1. And there's probability 0.5 that I stay at state 2. If I were to start from state 2, the probability that I go to 1 in 1 time step is this transition that has probability 0.2, and the other 0.8.

OK. So the calculation now becomes more interesting, if we want to calculate the next term. How likely is that at time 2, I find myself at state 1? In order to be here at state 1, this can happen in 2 ways. Either the first transition left me there, and the second transition is the same. So these correspond to this 0.5, that the first transition took me there, and the next transition was also of the same kind. That's one possibility. But there's another scenario. In order to be at state 1 at time 2 -- this can also happen this way. So that's the event that, after 1 transition, I got there. And the next transition happened to be this one.

So this corresponds to 0.5 times 0.2. It corresponds to taking the 1-step transition probability of getting there, times the probability that from state 2 I move to state 1, which in this case, is 0.2.

So basically we take this number, multiplied with 0.2, and then add those 2 numbers. And after you add them, you get 0.35. And similarly here, you're going to get 0.65.

And now to continue with the recursion, we keep doing the same thing. We take this number times 0.5 plus this number times 0.2. Add them up, you get the next entry. Keep doing that, keep doing that, and eventually you will notice that the numbers start settling into a limiting value at $2/7$. And let's verify this. If this number is $2/7$, what is the next number going to be? The next number is going to be $2/7$ -- (not 2.7) -- it's going to be $2/7$. That's the probability that I find myself at that state, times 0.5 -- that's the next transition that takes me to state 1 -- plus $5/7$ -- that would be the remaining probability that I find myself in state 2 -- times $1/5$. And so that gives me, again, $2/7$.

So this calculation basically illustrates, if this number has become $2/7$, then the next number is also going to be $2/7$. And of course this number here is going to have to be $5/7$. And this one would have to be again, the same, $5/7$. So the probability that I find myself at state 1, after a long time has elapsed, settles into some steady state value. So that's an interesting phenomenon. We just make this observation.

Now we can also do the calculation about the probability, starting from state 2. And here, you do the calculations -- I'm not going to do them. But after you do them, you find this probability also settles to $2/7$ and this one also settles to $5/7$. So these numbers here are the same as those numbers. What's the difference between these?

This is the probability that I find myself at state 1 given that I started at 1. This is the probability that I find myself at state 1 given that I started at state 2. These probabilities are the same, no matter where I started from. So this numerical example sort of illustrates the idea that after the chain has run for a long time, what the state of the chain is, does not care about the initial state of the chain.

So if you start here, you know that you're going to stay here for some time, a few transitions, because this probability is kind of small. So the initial state does that's tell you something. But in the very long run, transitions of this kind are going to happen. Transitions of that kind are going to happen. There's a lot of randomness that comes in, and that randomness washes out any information that could come from the initial state of the system.

We describe this situation by saying that the Markov chain eventually enters a steady state. Where a steady state, what does it mean it? Does it mean the state itself becomes steady and stops at one place? No, the state of the chain keeps jumping forever. The state of the chain will keep making transitions, will keep going back and forth between 1 and 2. So the state itself, the X_n , does not become steady in any sense.

What becomes steady are the probabilities that describe X_n . That is, after a long time elapses, the probability that you find yourself at state 1 becomes a constant $2/7$, and the probability that you find yourself in state 2 becomes a constant. So jumps will keep happening, but at any given time, if you ask what's the probability that right now I am at state 1, the answer is going to be $2/7$.

Incidentally, do the numbers sort of makes sense? Why is this number bigger than that number? Well, this state is a little more sticky than that state. Once you enter here, it's kind of harder to get out. So when you enter here, you spend a lot of time here. This one is easier to get out, because the probability is 0.5, so when you enter there, you tend to get out faster. So you keep moving from one to the other, but you tend to spend more time on that state, and this is reflected in this probability being bigger than that one. So no matter where you start, there's 5/7 probability of being here, 2/7 probability being there.

So there were some really nice things that happened in this example. The question is, whether things are always as nice for general Markov chains. The two nice things that happened where the following-- as we keep doing this calculation, this number settles to something. The limit exists. The other thing that happens is that this number is the same as that number, which means that the initial state does not matter. Is this always the case? Is it always the case that as n goes to infinity, the transition probabilities converge to something?

And if they do converge to something, is it the case that the limit is not affected by the initial state i at which the chain started? So mathematically speaking, the question we are raising is whether $R_{ij}(n)$ converges to something. And whether that something to which it converges to has only to do with j . It's the probability that you find yourself at state j , and that probability doesn't care about the initial state. So it's the question of whether the initial state gets forgotten in the long run.

So the answer is that usually, or for nice chains, both of these things will be true. You get the limit which does not depend on the initial state. But if your chain has some peculiar or unique structure, this might not happen. So let's think first about the issue of convergence. So convergence, as n goes to infinity at a steady value, really means the following. If I tell you a lot of time has passed, then you tell me, OK, the state of the probabilities are equal to that value without having to consult your clock. If you don't have convergence, it means that R_{ij} can keep going up and down, without settling to something. So in order for you to tell me the value of R_{ij} , you need to consult your clock to check if, right now, it's up or is it down.

So there's some kind of periodic behavior that you might get when you do not get convergence, and this example here illustrates it. So what's happened in this example? Starting from state 2, next time you go here, or there, with probability half. And then next time, no matter where you are, you move back to state 2. So this chain has some randomness, but the randomness is kind of limited type. You go out, you come in. You go out, you come in. So there's a periodic pattern that gets repeated. It means that if you start at state 2 after an even number of steps, you are certain to be back at state 2. So this probability here is 1.

On the other hand, if the number of transitions is odd, there's no way that you can be at your initial state. If you start here, at even times you would be here, at odd times you would be there or there. So this probability is 0. As n goes to infinity, these probabilities, the n -step transition probability does not converge to anything. It keeps alternating between 0 and 1. So convergence fails.

This is the main mechanism by which convergence can fail if your chain has a periodic structure. And we're going to discuss next time that, if periodicity absent, then we don't have an issue with convergence. The second question if we have convergence, whether the initial state matters or not. In the previous chain, where you could keep going back and forth between states 1 and 2 numerically, one finds that the initial state does not matter. But you can think of situations where the initial state does matter. Look at this chain here.

If you start at state 1, you stay at state 1 forever. There's no way to escape. So this means that $R_{11}(n)$ is 1 for all n . If you start at state 3, you will be moving between stage 3 and 4, but there's no way to go in that direction, so there's no way that you go to state 1. And for that reason, R_{31} is 0 for all n .

OK So this is a case where the initial state matters. R_{11} goes to a limit, as n goes to infinity, because it's constant. It's always 1 so the limit is 1. R_{31} also has a limit. It's 0 for all times. So these are the long term probabilities of finding yourself at state 1. But those long-term probabilities are affected by where you started. If you start here, you're sure that's, in the long term, you'll be here. If you start here, you're sure that, in the long term, you will not be there. So the initial state does matter here.

And this is a situation where certain states are not accessible from certain other states, so it has something to do with the graph structure of our Markov chain. Finally let's answer this question here, at least for large n 's. What do you think is going to happen in the long term if you start at state 2? If you start at state 2, you may stay at state 2 for a random amount of time, but eventually this transition will happen, or that transition would happen. Because of the symmetry, you are as likely to escape from state 2 in this direction, or in that direction, so there's probability 1/2 that, when the transition happens, the transition happens in that direction.

So for large N , you're certain that the transition does happen. And given that the transition has happened, it has probability 1/2 that it has gone that particular way. So clearly here, you see that the probability of finding yourself in a particular state is very much affected by where you started from. So what we want to do next is to abstract from these two examples and describe the general structural properties that have to do with periodicity, and that have to do with what happened here with certain states, not being accessible from the others.

We're going to leave periodicity for next time. But let's talk about the second kind of phenomenon that we have. So here, what we're going to do is to classify the states in a transition diagram into two types, recurrent and transient. So a state is said to be recurrent if the following is true. If you start from the state i , you can go to some places, but no matter where you go, there is a way of coming back. So what's an example for the recurrent state? This one. Starting from here, you can go elsewhere. You can go to state 7. You can go to state 6. That's all where you can go to. But no matter where you go, there is a path that can take you back there.

So no matter where you go, there is a chance, and there is a way for returning where you started. Those states we call recurrent. And by this, 8 is recurrent. All of these are recurrent. So this is recurrent, this is recurrent. And this state 5 is also recurrent. You cannot go anywhere from 5 except to 5 itself. Wherever you can go, you can go back to where you start. So this is recurrent.

If it is not the recurrent, we say that it is transient. So what does transient mean? You need to take this definition, and reverse it. Transient means that, starting from i , there is a place to which you could go, and from which you cannot return. If it's recurrent, anywhere you go, you can always come back. Transient means there are places where you can go from which you cannot come back.

So state 1 is recurrent - because starting from here, there's a possibility that you get there, and then there's no way back. State 4 is recurrent, starting from 4, there's somewhere you can go and - sorry, transient, correct. State 4 is transient starting from here, there are places where you could go, and from which you cannot come back. And in this particular diagram, all these 4 states are transients.

Now if the state is transient, it means that there is a way to go somewhere where you're going to get stuck and not to be able to come. As long as your state keeps circulating around here, eventually one of these transitions is going to happen, and once that happens, then there's no way that you can come back. So that transient state will be visited only a finite number of times. You will not be able to return to it. And in the long run, you're certain that you're going to get out of the transient states, and get to some class of recurrent states, and get stuck forever.

So, let's see, in this diagram, if I start here, could I stay in this lump of states forever? Well as long as I'm staying in this type of states, I would keep visiting states 1 and 2. Each time that I visit state 2, there's going to be positive probability that I escape. So in the long run, if I were to stay here, I would visit state 2 an infinite number of times, and I would get infinite chances to escape. But if you have infinite chances to escape, eventually you will escape. So you are certain that with probability 1, starting from here, you're going to move either to those states, or to those states.

So starting from transient states, you only stay at the transient states for random but finite amount of time. And after that happens, you end up in a class of recurrent states. And when I say class, what they mean is that, in this picture, I divide the recurrent states into 2 classes, or categories. What's special about them? These states are recurrent. These states are recurrent. But there's no communication between the 2. If you start here, you're stuck here. If you start here, you are stuck there.

And this is a case where the initial state does matter, because if you start here, you get stuck here. You start here, you get stuck there. So depending on the initial state, that's going to affect the long term behavior of your chain. So the guess you can make at this point is that, for the initial state to not matter, we should not have multiple recurrent classes. We should have only 1. But we're going to get back to this point next time.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013

Transcript – Recitation: Markov Chain Practice 1

Hi, everyone. Today, I'm going to talk about Markov Chain Practice number one. Before we start, let's first take a look at this Markov chain. This Markov chain has six states. In this problem, we always assume the process starts from state S0. On the first trial, the process can either make a transition from S0 to S1 with probability 1/3 or from S0 to S3 with probability 1/3 third or from S0 to S5 with probability 1/3.

If on the first trial, the process makes the transition from S0 to S1 or from S0 to S5, it will always be stuck in either S1 or S5 forever, because both of the states S1 and S5 have a self-transition probability of one. On the other hand, if on the first trial, the process makes the transition from S0 to S3, it can then either transition to the left or transition to the right or make self-transition back to the state S3.

If the process ever enters the left of the chain, it will never be able to come to the right. On the other hand, if the process ever enters the right of the chain, it would never be able to go to the left. For part A of the problem, we have to calculate the probability that the process enter S2 for the first time at the case trial.

First, notice that it would take at least two trials for the process to make a transition from S0 to S2. Therefore, for k equal to 1, the probability of a_k is simply equal to 0. For k equal to 1, probability of a_1 is equal to 0. Then for k equal to 2, 3 and on, the probability that the process enters S2 for the first time at a case trial is equivalent to the probability that the process first makes a transition from S0 to S3 and then stays in S3 for the next two k minus 2 trials and finally makes a transition from S3 to S2 on the kth trial.

So let's write this out. For k equal to 2, 3, and on, the probability of a_k is equal to the probability that the process first makes transition from S0 to S3 on the first trial, which is probability 03, times the probability that the process makes self-transition for the next k minus 2 trials, which is probability 33 to the power of k minus 2, and finally makes a transition from S3 to S2 on the kth trial, which is p_{32} . And this gives us 1/3 times 1/4 to the power of k minus 2 times 1/4, which is equal to 1/3 times 1/4 to the power of k minus--

For part B of the problem, we have to calculate the probability that the process never enters as four. This event can happen in three ways. The first way is that the process makes a transition from S0 to S1 on the first trial and be stuck in S1 forever. The second way that the process makes a transition from S0 to S5 on the first trial and be stuck at S5 forever. The third way is that the process makes a transition from S0 to S3 on the first trial and then it makes a transition from S3 to S2 on the next state change so that it would never be able to go to S4.

Therefore, the probability of B is equal to the sum of probabilities of this three events. So the probability of B is equal to the probability that the process makes a transition from S0 to S1 on the first trial, which is 1/3, plus the probability that the process makes a transition from S0 to S5

on the first trial, which is also $1/3$, plus the probability that the process makes a transition from S_0 to S_3 on the first trial times the probability that the process then makes a transition from S_3 to S_2 on the next state change.

So transition to S_2 , given that the processes are already in state S_3 and there's a state change. Let's take a look at this conditional probability. The condition that the processes are already in state S_3 and there's a state change imply two possible events, which are the transition from S_3 to S_2 and the transition from S_3 to S_4 . Therefore, we can write this conditional probability as the conditional probability of transition from S_3 to S_2 , given that another event, S_3 to S_2 or S_3 to S_4 has happened.

And this is simply equal to the proportion of p_{32} and $p_{32} + p_{34}$, which is equal to $1/4$ over $1/4 + 1/2$, which is equal to $1/3$. Therefore, the probability of B is equal to $1/3$ plus $1/3$ plus $1/3$ times the $1/3$ here, which is equal to $7/9$. For part C of the problem, we have to calculate the probability that the process enters S_2 and leaves S_2 on the next trial.

This probability can be written as the product of two probabilities-- the probability that the process enters S_2 and the probability that it leaves S_2 on the next trial, given it's already in S_2 . Let's first look at the probability that the process enters S_2 . Using a similar approach as part B, we know that the probability the process ever enters S_2 is equal to the probability of the event that the process first makes a transition from S_0 to S_3 on the first trial and then makes a transition from S_3 to S_2 on the next state change.

So the probability that the process enters S_2 is equal to the probability that it first makes a transition from S_0 to S_3 on the first trial, which is P_{03} , times the probability that it makes a transition to S_2 , given that it's already in S_3 and there is a state change. We have already calculated this conditional probability in part B. Let's then look at the second probability term, the probability that the process leaves S_2 on the next trial, given that it's already in S_2 .

So given that the process is already in S_2 , it can take two transitions. It can either transition from S_2 to S_1 or make a self-transition from S_2 back to S_2 . Therefore, this conditional probability that it leaves S_2 on the next trial, given that it was already in S_2 is simply equal to the transition probability from S_2 to S_1 , which is P_{21} . Therefore, this is equal to P_{03} , which is $1/3$, times $1/3$ from the result from part B times P_{21} , which is $1/2$, and gives us $1/18$.

For part D of the problem, we have to calculate the probability that the process enters S_1 for the first time on the third trial. So if you take a look at this Markov chain, you'll notice that the only way for this event to happen is when a process first makes a transition from S_0 to S_3 on the first trial and from S_3 to S_2 on the second trial and from S_2 to S_1 on the third trial. Therefore, the probability of D is equal to the probability of the event that the process makes a transition from S_0 to S_3 on the first trial and from S_3 to S_2 on the second trial and finally from S_2 to S_1 on the third trial.

So this is equal to P_{03} times P_{32} times P_{21} , which is equal to $1/3$ times $1/4$ times $1/2$, which is equal to $1/24$. For part E of the problem, we have to calculate the probability that the process is in S_3 immediately after the n th trial. If you take a look at this Markov chain, you'll notice that if

on the first trial, the process makes a transition from S0 to S1 or from S0 to S5, it will never be able to go to S3.

On the other hand, if on the first trial, the process makes a transition from S0 to S3 and if it leaves S3 at some point, it will never be able to come back to S3. Therefore, in order for the process to be S3 immediately after the nth trial, we will need the process to first make transition from S0 to S3 on the first trial and then stay in S3 for the next n minus 1 trials. Therefore, the probability of the event e is simply equal to the probability of this event, which is equal to P03 times P33 to the power n minus 1, which is equal to 1/3 times 1/4 to the power of n minus 1.

And this concludes our practice on Markov chain today.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013

Transcript – Recitation: Setting Up a Markov Chain

Hi. In this problem, we're going to practice setting up a Markov chain by going fishing in this lake, which has n fish in it, some of which are green. And the rest of the fish are blue. So, what we do is, every day we go to this lake, and we catch exactly 1 fish. And all the fish are equally likely to be the 1 that's caught.

Now, if we catch a green fish, we paint it blue, and we throw back into the lake. And if we catch a blue fish, we just keep it blue, and we also throw it back.

Now, what we're interested in modeling is, how does this lake evolve over time? And specifically what we're interested in is the number of green fish that are left in the lake. So, let's let G_i be the event that there are i green fish left in the lake. And we want to know, how does G_i evolve over time?

Now, one thing that we've learned that we can use to model this is a Markov chain. But before we can use it, we need to make sure that this actually satisfies the Markov property. Now, recall that the Markov property essentially says that, given the current state of the system, that's all you need in order to predict the future states. So, any past history of the previous states that it was in, that's all irrelevant. All you need is the current state.

Now, in the context of this particular problem, what that means is that if I tell you that there are 10 green fish left, that's all the information you need in order to predict how many fish there will be tomorrow. So, why is that?

Well, it's because what influences the number of green fish that are left? What influences it is which fish you catch because, depending on which fish you catch, you may paint the green fish blue, in which case the number of green fish decrease. But what affects which fish you catch?

Well, that probability is dictated solely based on just the number of green fish in the lake right now, today. So, it doesn't matter that there were 20 fish yesterday. All that matters is how many green fish there are in the lake today. And so, because of that argument, the number of green fish-- this does satisfy the Markov property, so we can use this and model it as a Markov chain.

So, like we alluded to just now, the key dynamic that we need to look at is, how does the number of green fish change? And if we look at it, we notice that after each day, the number of green fish can only have two possible transitions.

One possible transition is that it goes down by exactly 1, which happens if you happen to catch a green fish and paint it blue. So, that green fish is no longer green, so the number of green fish goes down by 1. The other possible transition is that G_i doesn't change because you caught a blue fish that day. So, all the green fish are still green.

So, now given that, let's see if we can come up with a Markov chain. So, the first thing we've done is we've written down all the different states, right? So, this represents the number of green fish left in the lake. So, there could be 0 green fish left, 1 green fish, all the way through n , which means that every single fish in the lake is green.

Now, we have the states. What we need to do now is to fill in the transition probabilities, which are the P_{ij} 's. And remember, the P_{ij} is the probability of transitioning from state i to state j in the next transition. So, what that means in this context is, what's the probability that there will be j green fish tomorrow given that there are i green fish today?

Now, if we go back to our earlier argument, we see that for any given i , you can only transition to two possible j 's. One of them is you stay at i because the number of green fish doesn't change because you caught a blue fish. And the other is that you'd go from i to $i - 1$. The number of green fish decreases by 1.

Now, what we need to do now is fill in what those probabilities are. So, if j equals i , meaning that the number of green fish doesn't change, well, what's the probability that you have the same number of green fish tomorrow as you do today? Well, if you have i green fish today, that happens if you catch 1 of the $n - i$ blue fish. So, what's the probability of catching one of the $n - i$ blue fish? Well, it's $n - i$ over n .

Now, the other possible transition is you go from i to j equals $i - 1$, so i goes down by 1. And that happens when you catch a green fish. So, given that there are i green fish, what's the probability that you catch 1 of those? Well, it's going to be i/n .

And finally, every other transition has 0 probability.

All right. So, now we can add those transitions on to our Markov chain. So, for example, we have these. So, let's look at this general case i . So, if you're state i , you have i green fish left. You will transition to $i - 1$ green fish left if that day you caught a green fish. And we said that that probability is i/n .

And the self transition probability is you caught a blue fish that day, so you still stay at i green fish. And that probability, we said, was $n - i$ over n .

All right. Now, it's helpful to verify that this formula works by looking at some cases where it's intuitive to calculate what these probabilities should be. So, let's look at state n . That is the state where every single fish in the lake is green. So, if every single fish in the lake is green, then no matter what fish you catch, it's going to be green. And you're going to paint it blue and return it, so you're guaranteed to go down to $n - 1$ green fish.

And so, this transition probability down to $n - 1$ is guaranteed to be 1. And so, the self transition probability has to be 0. Now, let's go back to our formula and verify that actually gives us the right value.

So, if i is n , then there's only these transition probabilities. So, if i is n , then the transition probability to j , for j is also n , is n minus n over n , which is 0. And that's exactly what we said. We argued that the self transition probability should be 0.

And also, if i is n , the probability of transitioning to n minus 1 should be n over n , which is 1. And that's exactly what we argued here.

So, it seems like these transition probabilities do make sense. And if we wanted to, we could fill in the rest of these. So, for example, this would be $2/n$, $1/n$, n minus 1 over n , n minus 2 over n .

And now, let's also consider the case of state 0, which means that every single fish is blue. There are 0 green fish left. Well, if that's the case, then what's the probability of staying at 0?

Well, that's n minus 0 over n is 1, all right? So, the self transition probability is 1. And that makes sense because if you have 0 green fish, there's no way to generate more green fish because you don't paint blue fish green. And so, you're going to stay at 0 green fish forever.

All right. So, we've characterized the entire Markov chain now. And so, now let's just answer some simple questions about this. So, the problem asks us to identify, what are the recurrent and transient states?

So, remember that recurrent state means that if you start out at that state, no matter where you go, what other states you end up at, there is some positive probability path that will take you back to your original state. And if you're not recurrent, then you're transient, which means that if you're transient, if you start out at the transient state, there is some other state that you can go to, from which there's no way to come back to the original transient state.

All right. So, now let's look at this and see which states are recurrent and which are transient. And we can fill this in more.

And if we look at it, let's look at state n . Well, we're guaranteed to go from state n to state n minus 1. And once we're in state n minus 1, there's no way for us to go back to state n because we can't generate more green fish. And so, n is transient.

And similarly, we can use the same argument to show that everything from 1 through n , all of these states, are transient for the same reason because there's no way to generate more green fish. And so, the chain can only stay at a given state or go down 1. And so, it always goes down. It can only go left, and it can never go right. So, once you leave a certain state, there's no way to come back.

And so, states 1 through n are all transient. And 0 the only recurrent state because, well, the only place you go from 0 is itself. So, you always stay at 0. And in fact, 0 is not only recurrent, it's absorbing because every single other state, no matter where you start out at, you will always end up at 0.

So, this was just an example of how to set up a Markov chain. You just think about the actual dynamics of what's going on and make sure that it satisfies the Markov property. Then, figure out what all the states are and calculate all the transition probabilities. And once you have that, you've specified your Markov chain.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

LECTURE 16

Markov Processes – I

- **Readings:** Sections 7.1–7.2

Lecture outline

- Checkout counter example
- Markov process definition
- n -step transition probabilities
- Classification of states

Checkout counter model

- Discrete time $n = 0, 1, \dots$
- Customer arrivals: Bernoulli(p)
 - geometric interarrival times
- Customer service times: geometric(q)
- “State” X_n : number of customers at time n



Finite state Markov chains

- X_n : state after n transitions
 - belongs to a finite set, e.g., $\{1, \dots, m\}$
 - X_0 is either given or random
- **Markov property/assumption:**
(given current state, the past does not matter)

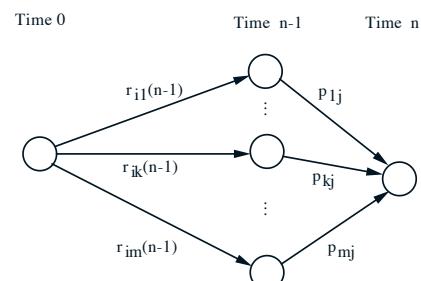
$$p_{ij} = P(X_{n+1} = j | X_n = i)$$

$$= P(X_{n+1} = j | X_n = i, X_{n-1}, \dots, X_0)$$
- Model specification:
 - identify the possible states
 - identify the possible transitions
 - identify the transition probabilities

n -step transition probabilities

- State occupancy probabilities, given initial state i :

$$r_{ij}(n) = P(X_n = j | X_0 = i)$$



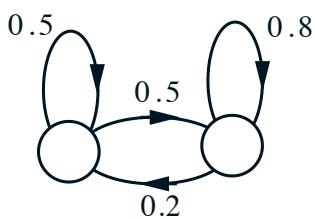
- Key recursion:

$$r_{ij}(n) = \sum_{k=1}^m r_{ik}(n-1) p_{kj}$$

- With random initial state:

$$P(X_n = j) = \sum_{i=1}^m P(X_0 = i) r_{ij}(n)$$

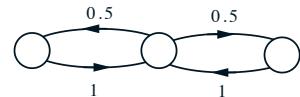
Example



	$n = 0$	$n = 1$	$n = 2$	$n = 100$	$n = 101$
$r_{11}(n)$					
$r_{12}(n)$					
$r_{21}(n)$					
$r_{22}(n)$					

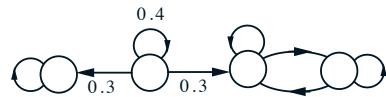
Generic convergence questions:

- Does $r_{ij}(n)$ converge to something?



n odd: $r_{22}(n) =$ n even: $r_{22}(n) =$

- Does the limit depend on initial state?



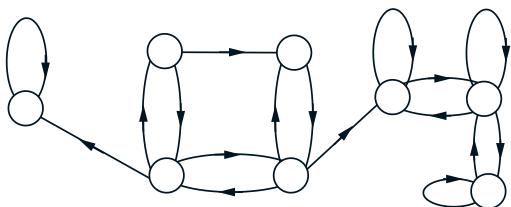
$r_{11}(n) =$

$r_{31}(n) =$

$r_{21}(n) =$

Recurrent and transient states

- State i is **recurrent** if:
starting from i ,
and from wherever you can go,
there is a way of returning to i
- If not recurrent, called **transient**



- i transient:
 $P(X_n = i) \rightarrow 0$,
 i visited finite number of times

- **Recurrent class:**
collection of recurrent states that
“communicate” with each other
and with no other state

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

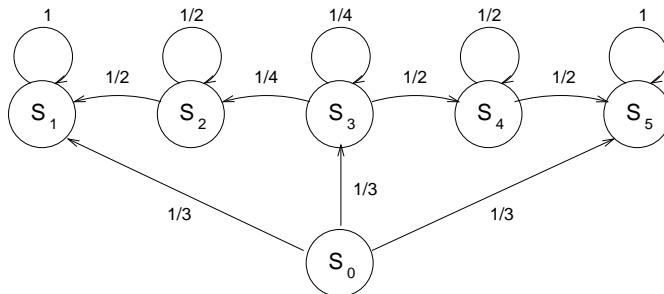
For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Recitation 18
November 9, 2010

1. There are n fish in a lake, some of which are green and the rest blue. Each day, Helen catches 1 fish. She is equally likely to catch any one of the n fish in the lake. She throws back all the fish, but paints each green fish blue before throwing it back in. Let G_i denote the event that there are i green fish left in the lake.
 - (a) Show how to model this fishing exercise as a Markov chain, where $\{G_i\}$ are the states. Explain why your model satisfies the Markov property.
 - (b) Find the transition probabilities $\{p_{ij}\}$.
 - (c) List the transient and the recurrent states.

Textbook problem removed due to copyright restrictions.
 Drake, Fundamentals of Applied Probability Theory, Problem 5.02.

3. Consider the following Markov chain, with states labelled from s_0, s_1, \dots, s_5 :



Given that the above process is in state s_0 just before the first trial, determine by inspection the probability that:

- (a) The process enters s_2 for the first time as the result of the k th trial.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Fall 2010)

- (b) The process never enters s_4 .
- (c) The process enters s_2 and then leaves s_2 on the next trial.
- (d) The process enters s_1 for the first time on the third trial.
- (e) The process is in state s_3 immediately after the n th trial.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Recitation 18: Solutions
November 9, 2010

1. a) The number of remaining green fish at time n completely determines all the relevant information of the system's entire history (relevant to predicting the future state.) Therefore it is immediate that the number of green fish is the state of the system and the process has the Markov property:

$$\mathbf{P}(X_{m+1} = j | X_m = i, X_{m-1} = i_{m-1}, \dots, X_1 = i_1) = \mathbf{P}(X_{m+1} = j | X_m = i).$$

- b) For $j > i$ clearly $p_{ij} = 0$, since a blue fish will never be painted green. For $0 \leq i, j \leq k$, we have the following:

$$p_{ij} = \mathbf{P}(i - j \text{ green fish are caught} | \text{current state } = i) = \begin{cases} \frac{n-i}{n} & j = i \\ \frac{i}{n} & j = i-1 \\ 0 & \text{otherwise} \end{cases}$$

- c) The state 0 is an absorbing state since there is a positive probability that the system will enter it, and once it does, it will remain there forever. Therefore the state with 0 green fish is the only recurrent state, and all other states are then transient.

Textbook problem removed due to copyright restrictions.

Drake, Fundamentals of Applied Probability Theory, Problem 5.02.

3. (a) Let A_k be the event that the process enters s_2 for first time on trial k . The only way to enter state s_2 for the first time on the k th trial is to enter state s_3 on the first trial, remain in s_3 for the next $k - 2$ trials, and finally enter s_2 on the last trial. Thus,

$$\mathbf{P}(A_k) = p_{03} \cdot p_{33}^{k-2} \cdot p_{32} = \left(\frac{1}{3}\right) \left(\frac{1}{4}\right)^{k-2} \left(\frac{1}{4}\right) = \frac{1}{3} \left(\frac{1}{4}\right)^{k-1} \quad \text{for } k = 2, 3, \dots$$

- (b) Let A be the event that the process never enters s_4 .

There are three possible ways for A to occur. The first two are if the first transition is either from s_0 to s_1 or s_0 to s_5 . This occurs with probability $\frac{2}{3}$. The other is if The first transition is from s_0 to s_3 , and that the next change of state *after* that is to the state s_2 . We know that the probability of going from s_0 to s_3 is $\frac{1}{3}$. Given this has occurred, and given a change of state occurs from state s_3 , we know that the probability that the state transitioned to is the state s_2 is simply $\frac{\frac{1}{4}}{\frac{1}{4} + \frac{1}{2}} = \frac{1}{3}$. Thus, the probability of transitioning from s_0 to s_3 and then eventually transitioning to s_2 is $\frac{1}{9}$. Thus, the probability of never entering s_4 is $\frac{2}{3} + \frac{1}{9} = \frac{7}{9}$.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
 (Fall 2010)

(c) $\mathbf{P}(\{\text{process enters } s_2 \text{ and then leaves } s_2 \text{ on next trial}\})$

$$\begin{aligned}
 &= \mathbf{P}(\{\text{process enters } s_2\})\mathbf{P}(\{\text{leaves } s_2 \text{ on next trial}\} | \{\text{in } s_2\}) \\
 &= \left[\sum_{k=2}^{\infty} \mathbf{P}(A_k) \right] \cdot \frac{1}{2} \\
 &= \left[\sum_{k=2}^{\infty} \frac{1}{3} \left(\frac{1}{4}\right)^{k-1} \right] \cdot \frac{1}{2} \\
 &= \frac{1}{6} \cdot \frac{\frac{1}{4}}{1 - \frac{1}{4}} \\
 &= \frac{1}{18}.
 \end{aligned}$$

(d) This event can only happen if the sequence of state transitions is as follows:

$$s_0 \longrightarrow s_3 \longrightarrow s_2 \longrightarrow s_1.$$

Thus, $\mathbf{P}(\{\text{process enters } s_1 \text{ for first time on third trial}\}) = p_{03} \cdot p_{32} \cdot p_{21} = \frac{1}{3} \cdot \frac{1}{4} \cdot \frac{1}{2} = \frac{1}{24}$.

(e) $\mathbf{P}(\{\text{process in } s_3 \text{ immediately after the } N\text{th trial}\})$

$$\begin{aligned}
 &= \mathbf{P}(\{\text{moves to } s_3 \text{ in first trial and stays in } s_3 \text{ for next } N-1 \text{ trials}\}) \\
 &= \frac{1}{3} \left(\frac{1}{4}\right)^{n-1} \quad \text{for } n = 1, 2, 3, \dots
 \end{aligned}$$

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Tutorial/Recitation 9
November 12, 2010

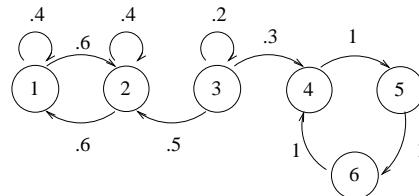
1. Problem 7.13, page 385 in textbook.

The times between successive customer arrivals at a facility are independent and identically distributed random variables with the following PMF:

$$p(k) = \begin{cases} 0.2, & k = 1 \\ 0.3, & k = 3 \\ 0.5, & k = 4 \\ 0, & \text{otherwise} \end{cases}$$

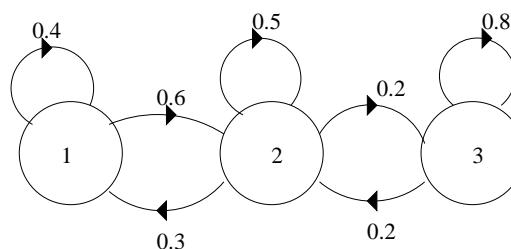
Construct a four-state Markov chain model that describes the arrival process. In this model, one of the states should correspond to the times when an arrival occurs.

2. The Markov chain shown below is in state 3 immediately before the first trial.



- (a) Indicate which states, if any, are recurrent, transient, and periodic.
 - (b) Find the probability that the process is in state 3 after n trials.
 - (c) Find the expected number of trials up to and including the trial on which the process leaves state 3.
 - (d) Find the probability that the process never enters state 1.
 - (e) Find the probability that the process is in state 4 after 10 trials.
 - (f) Given that the process is in state 4 after 10 trials, find the probability that the process was in state 4 after the first trial.
3. Problem 7.13, page 385 in textbook.

Consider the Markov chain below. Let us refer to a transition that results in a state with a higher (respectively, lower) index as a birth (respectively, death). Calculate the following quantities, assuming that when we start observing the chain, it is already in steady-state.



- (a) For each state i , the probability that the current state is i .

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Fall 2010)

- (b) The probability that the first transition we observe is a birth.
- (c) The probability that the first change of state we observe is a birth.
- (d) The conditional probability that the process was in state 2 before the first transition that we observe, given that this transition was a birth.
- (e) The conditional probability that the process was in state 2 before the first change of state that we observe, given that this change of state was a birth.
- (f) The conditional probability that the first observed transition is a birth given that it resulted in a change of state.
- (g) The conditional probability that the first observed transition leads to state 2, given that it resulted in a change of state.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.04SC Probabilistic Systems Analysis and Applied Probability

Fall 2010

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Tutorial/Recitation 9: Solutions

1. Problem 7.1, page 380 in textbook. See online solutions.
2. (a) Recurrent: 1, 2, 4, 5 , 6; Transient: 3; Periodic: 4,5,6.
(b) 0.2^n
(c) This is a geometric random variable with parameter $p = 0.5 + 0.3$. Hence, the expected number of trials up to and includ ing the trial on which the process leaves state 3 is $\mathbf{E}[X] = 1/p = 5/4$.
(d) $3/8$
(e) $\mathbf{P}(A) = 0.3 + 0.2^30.3 + 0.2^60.3 + 0.2^90.3 = 0.3024$.
(f) $0.3/\mathbf{P}(A) = 0.992$.
3. Problem 7.13, page 385 in textbook. See online solutions.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013 Transcript – Lecture 17

The following content is provided under a Creative Commons license. Your support will help MIT OpenCourseWare continue to offer high-quality, educational resources for free. To make a donation or view additional materials from hundreds of MIT courses, visit MIT OpenCourseWare at ocw.mit.edu.

PROFESSOR: All right. So today, we're going to start by taking stock of what we discussed last time, review the definition of Markov chains. And then most of the lecture, we're going to concentrate on their steady-state behavior. Meaning, we're going to look at what does a Markov chain do if it has run for a long time. What can we say about the probabilities of the different states?

So what I would like to repeat is a statement I made last time that Markov chains is a very, very useful class of models. Pretty much anything in the real world can be approximately modeled by a Markov chain provided that you set your states in the proper way. So we're going to see some examples. You're going to see more examples in the problems you're going to do in homework and recitation.

On the other hand, we're not going to go too deep into examples. Rather, we're going to develop the general methodology. OK.

All right. Markov models can be pretty general. They can run in continuous or discrete time. They can have continuous or discrete state spaces. In this class, we're going to stick just to the case where the state space is discrete and time is discrete because this is the simplest case. And also, it's the one where you build your intuition before going to more general cases perhaps in other classes.

So the state is discrete and finite. There's a finite number of states. At any point in time, the process is sitting on one of those states. Time is discrete, so at each unit of time, somebody whistles and then the state jumps. And when it jumps, it can either land in the same place, or it can land somewhere else. And the evolution of the process is described by transition probabilities.

P_{ij} is the probability that the next state is j given that the current state is i . And the most important property that the Markov chain has, the definition of a Markov chain or Markov process, is that this probability, P_{ij} , is the same every time that you land at state i -- no matter how you got there and also no matter what time it is.

So the model we have is time homogeneous, which basically means that those transition probabilities are the same at every time. So the model is time invariant in that sense. So we're interested in what the chain or the process is going to do in the longer run. So we're interested,

let's say, in the probability that starting at a certain state, n times steps later, we find ourselves at some particular state j .

Fortunately, we can calculate those probabilities recursively. Of course, at the first time 1, the probability of being 1 time later at state j given that we are right now at state i , by definition, this is just the transition probabilities. So by knowing these, we can start a recursion that tells us the transition probabilities for more than n steps.

This recursion, it's a formula. It's always true. You can copy it or memorize it. But there is a big idea behind that formula that you should keep in mind. And basically, the divide and conquer idea. It's an application of the total probability law. So let's fix i . The probability that you find yourself at state j , you break it up into the probabilities of the different ways that you can get to state j .

What are those different ways? The different ways are the different states k at which you might find yourself the previous time. So with some probability, with this probability, you find yourself at state k the previous time. And then with probability P_{kj} , you make a transition to state j . So this is a possible scenario that takes you to state j after n transitions. And by summing over all the k 's, then we have considered all the possible scenarios.

Now, before we move to the more serious stuff, let's do a little bit of warm up to get a handle on how we use transition probabilities to calculate more general probabilities, then talk about some structural properties of Markov chains, and then eventually get to the main business of today, which is a steady-state behavior.

So somebody gives you this chain, and our convention is that those arcs that are not shown here corresponds to 0 probabilities. And each one of the arcs that's shown has a non-zero probability, and somebody gives it to us. Suppose that the chain starts at state 1. We want to calculate the probability that it follows this particular path. That is, it goes to 2, then to 6, then to 7. How do we calculate the probability of a particular trajectory?

Well, this is the probability-- so it's the probability of the trajectory from 1 that you go to 2, then to 6, then to 7. So the probability of this trajectory is we use the multiplication rule. The probability of several things happening is the probability that the first thing happens, which is a transition from 1 to 2. And then given that we are at state 2, we multiply with a conditional probability that the next event happens. That is, that X_2 is equal to 6 given that right now, we are at state 1. And that conditional probability is just P_{26} .

And notice that this conditional probability applies no matter how we got to state 2. This is the Markov assumption. So we don't care about the fact that we came in in a particular way. Given that we came in here, this probability P_{26} , that the next transition takes us to 6. And then given that all that stuff happened, so given that right now, we are at state 6, we need to multiply with a conditional probability that the next transition takes us to state 7. And this is just the P_{67} .

So to find the probability of following a specific trajectory, you just multiply the transition probabilities along the particular trajectory. Now, if you want to calculate something else, such

as for example, the probability that 4 time steps later, I find myself at state 7 given that they started, let's say, at this state. How do you calculate this probability?

One way is to use the recursion for the Rij's that we know that it is always valid. But for short and simple examples, and with a small time horizon, perhaps you can do this in a brute force way. What would be the brute force way? This is the event that 4 time steps later, I find myself at state 7. This event can happen in various ways. So we can take stock of all the different ways, and write down their probabilities.

So starting from 2. One possibility is to follow this trajectory, 1 transition, 2 transitions, 3 transitions, 4 transitions. And that takes me to state 7. What's the probability of this trajectory? It's P_{26} times P_{67} times P_{76} and then times P_{67} . So this is a probability of a particular trajectory that takes you to state 7 after 4 time steps.

But there's other trajectories as well. What could be it? I might start from state 2, go to state 6, stay at state 6, stay at state 6 once more. And then from state 6, go to state 7. And so there must be one more. What's the other one? I guess I could go 1, 2, 6, 7.

OK. That's the other trajectory. Plus P_{21} times P_{12} times P_{26} and times P_{67} . So the transition probability, the overall probability of finding ourselves at state 7, is broken down as the sum of the probabilities of all the different ways that I can get to state 7 in exactly 4 steps.

So we could always do that without knowing much about Markov chains or the general formula for the Rij's that we had. What's the trouble with this procedure? The trouble with this procedure is that the number of possible trajectories becomes quite large if this index is a little bigger. If this 4 was 100, and you ask how many different trajectories of length 100 are there to take me from here to there, that number of trajectories would be huge. It grows exponentially with the time horizon. And this kind of calculation would be impossible.

The basic equation, the recursion that have for the Rij's is basically a clever way of organizing this computation so that the amount of computation that you do is not exponential in the time horizon. Rather, it's sort of linear with the time horizon. For each time step you need in the time horizon, you just keep repeating the same iteration over and over.

OK. Now, the other thing that we discussed last time, briefly, was a classification of the different states of the Markov chain in two different types. A Markov chain, in general, has states that are recurrent, which means that from a recurrent state, I can go somewhere else. But from that somewhere else, there's always some way of coming back. So if you have a chain of this form, no matter where you go, no matter where you start, you can always come back where you started. States of this kind are called recurrent.

On the other hand, if you have a few states all this kind, a transition of this type, then these states are transient in the sense that from those states, it's possible to go somewhere else from which place there's no way to come back where you started. The general structure of a Markov chain is basically a collection of transient states. You're certain that you are going to leave the transient states eventually.

And after you leave the transient states, you enter into a class of states in which you are trapped. You are trapped if you get inside here. You are trapped if you get inside there. This is a recurrent class of states.

From any state, you can get to any other state within this class. That's another recurrent class. From any state inside here, you can get anywhere else inside that class. But these 2 classes, you do not communicate. If you start here, there's no way to get there.

If you have 2 recurrent classes, then it's clear that the initial conditions of your Markov chain matter in the long run. If you start here, you will be stuck inside here for the long run and similarly about here. So the initial conditions do make a difference. On the other hand, if this class was not here and you only had that class, what would happen to the chain?

Let's say you start here. You move around. At some point, you make that transition. You get stuck in here. And inside here, you keep circulating, because of the randomness, you keep visiting all states over and over. And hopefully or possibly, in the long run, it doesn't matter exactly what time it is or where you started, but the probability of being at that particular state is the same no matter what the initial condition was.

So with a single recurrent class, we hope that the initial conditions do not matter. With 2 or more recurrent classes, initial conditions will definitely matter. So how many recurrent classes we have is something that has to do with the long-term behavior of the chain and the extent to which initial conditions matter.

Another way that initial conditions may matter is if a chain has a periodic structure. There are many ways of defining periodicity. The one that I find sort of the most intuitive and with the least amount of mathematical symbols is the following. The state space of a chain is said to be periodic if you can lump the states into a number of clusters called d clusters or groups. And the transition diagram has the property that from a cluster, you always make a transition into the next cluster.

So here d is equal to 2. We have two subsets of the state space. Whenever we're here, next time we'll be there. Whenever we're here, next time we will be there. So this chain has a periodic structure. There may be still some randomness. When I jump from here to here, the state to which I jump may be random, but I'm sure that I'm going to be inside here. And then next time, I will be sure that I'm inside here.

This would be a structure of a diagram in which we have a period of 3. If you start in this lump, you know that the next time, you would be in a state inside here. Next time, you'll be in a state inside here, and so on. So these chains certainly have a periodic structure. And that periodicity gets maintained.

If I start, let's say, at this lump, at even times, I'm sure I'm here. At odd times, I'm sure I am here. So the exact time does matter in determining the probabilities of the different states. And in particular, the probability of being at the particular state cannot convert to a state value.

The probability of being at the state inside here is going to be 0 for all times. In general, it's going to be some positive number for even times. So it goes 0 positive, zero, positive, 0 positive. Doesn't settle to anything. So when we have periodicity, we do not expect the states probabilities to converge to something, but rather, we expect them to oscillate.

Now, how can we tell whether a Markov chain is periodic or not? There are systematic ways of doing it, but usually with the types of examples we see in this class, we just eyeball the chain, and we tell whether it's periodic or not. So is this chain down here, is it the periodic one or not? How many people think it's periodic?

No one. One. How many people think it's not periodic? OK. Not periodic? Let's see. Let me do some drawing here. OK. Is it periodic? It is. From a red state, you can only get to a white state. And from a white state, you can only get to a red state.

So this chain, even though it's not apparent from the picture, actually has this structure. We can group the states into red states and white states. And from reds, we always go to a white, and from a white, we always go to a red. So this tells you that sometimes eyeballing is not as easy. If you have lots and lots of states, you might have some trouble doing this exercise.

On the other hand, something very useful to know. Sometimes it's extremely easy to tell that the chain is not periodic. What's that case? Suppose that your chain has a self-transition somewhere. Then automatically, you know that your chain is not periodic.

So remember, the definition of periodicity requires that if you are in a certain group of states, next time, you will be in a different group. But if you have self-transitions, that property is not true. If you have a possible self-transition, it's possible that you stay inside your own group for the next time step.

So whenever you have a self-transition, this implies that the chain is not periodic. And usually that's the simplest and easy way that we can tell most of the time that the chain is not periodic.

So now, we come to the big topic of today, the central topic, which is the question about what does the chain do in the long run. The question we are asking and which we motivated last time by looking at an example. It's something that did happen in our example of last time. So we're asking whether this happens for every Markov chain.

We're asking the question whether the probability of being at state j at some time n settles to a steady-state value. Let's call it $\pi_{\text{sub } j}$. That these were asking whether this quantity has a limit as n goes to infinity, so that we can talk about the steady-state probability of state j .

And furthermore, we asked whether the steady-state probability of that state does not depend on the initial state. In other words, after the chain runs for a long, long time, it doesn't matter exactly what time it is, and it doesn't matter where the chain started from. You can tell me the probability that the state is a particular j is approximately the steady-state probability $\pi_{\text{sub } j}$. It doesn't matter exactly what time it is as long as you tell me that a lot of time has elapsed so that n is a big number.

So this is the question. We have seen examples, and we understand that this is not going to be the case always. For example, as I just discussed, if we have 2 recurrent classes, where we start does matter. The probability $\pi_i(j)$ of being in that state j is going to be 0 if we start here, but it would be something positive if we were to start in that lump. So the initial state does matter if we have multiple recurrent classes.

But if we have only a single class of recurrent states from each one of which you can get to any other one, then we don't have that problem. Then we expect initial conditions to be forgotten. So that's one condition that we need. And then the other condition that we need is that the chain is not periodic. If the chain is periodic, then these R_{ij} 's do not converge. They keep oscillating.

If we do not have periodicity, then there is hope that we will get the convergence that we need. It turns out this is the big theory of Markov chains-- the steady-state convergence theorem. It turns out that yes, the r_{ij} s do converge to a steady-state limit, which we call a steady-state probability as long as these two conditions are satisfied.

We're not going to prove this theorem. If you're really interested, the end of chapter exercises basically walk you through a proof of this result, but it's probably a little too much for doing it in this class.

What is the intuitive idea behind this theorem? Let's see. Let's think intuitively as to why the initial state doesn't matter. Think of two copies of the chain that starts at different initial states, and the state moves randomly. As the state moves randomly starting from the two initial states a random trajectory. as long as you have a single recurrent class at some point, and you don't have periodicity at some point, those states, those two trajectories, are going to collide. Just because there's enough randomness there. Even though we started from different places, the state is going to be the same.

After the state becomes the same, then the future of these trajectories, probabilistically, is the same because they both started at the same state. So this means that the initial conditions stopped having any influence. That's sort of the high-level idea of why the initial state gets forgotten. Even if you started at different initial states, at some time, you may find yourself to be in the same state as the other trajectory. And once that happens, your initial conditions cannot have any effect into the future.

All right. So let's see how we might calculate those steady-state probabilities. The way we calculate the steady-state probabilities is by taking this recursion, which is always true for the end-step transition probabilities, and take the limit of both sides. The limit of this side is the steady-state probability of state j , which is π_j . The limit of this side, we put the limit inside the summation. Now, as n goes to infinity, $n - 1$ also goes to infinity. So this R_{ik} is going to be the steady-state probability of state k starting from state i .

Now where we started doesn't matter. So this is just the steady-state probability of state k . So this term converges to that one, and this gives us an equation that's satisfied by the steady-state probabilities. Actually, it's not one equation. We get one equation for each one of the j 's. So if we

have 10 possible states, we're going to get the system of 10 linear equations. In the unknowns, $\pi(1)$ up to $\pi(10)$.

OK. 10 unknowns, 10 equations. You might think that we are in business. But actually, this system of equations is singular. 0 is a possible solution of this system. If you plug π equal to zero everywhere, the equations are satisfied. It does not have a unique solution, so maybe we need one more condition to get the uniquely solvable system of linear equations.

It turns out that this system of equations has a unique solution. If you impose an additional condition, which is pretty natural, the $\pi(j)$'s are the probabilities of the different states, so they should add to 1. So you want this one equation to the mix. And once you do that, then this system of equations is going to have a unique solution. And so we can find the steady-state probabilities of the Markov chain by just solving these linear equations, which is numerically straightforward.

Now, these equations are quite important. I mean, they're the central point in the Markov chain. They have a name. They're called the balance equations. And it's worth interpreting them in a somewhat different way. So intuitively, one can sometimes think of probabilities as frequencies. For example, if I toss an unbiased coin, probability 1/2 of heads, you could also say that if I keep flipping that coin, in the long run, 1/2 of the time, I'm going to see heads.

Similarly, let's try an interpretation of this $\pi(j)$, the steady-state probability, the long-term probability of finding myself at state j . Let's try to interpret it as the frequency with which I find myself at state j if I run a very, very long trajectory over that Markov chain. So the trajectory moves around, visits states. It visits the different states with different frequencies. And let's think of the probability that you are at a certain state as being sort of the same as the frequency of visiting that state.

This turns out to be a correct statement. If you were more rigorous, you would have to prove it. But it's an interpretation which is valid and which gives us a lot of intuition about what these equations are saying. So let's think as follows. Let's focus on a particular state j , and think of transitions into the state j versus transitions out of the state j , or transitions into j versus transitions starting from j . So transition starting from j includes a self-transition.

Ok. So how often do we get a transition, if we interpret the $\pi(j)$'s as frequencies, how often do we get a transition into j ? Here's how we think about it. A fraction $\pi(1)$ of the time, we're going to be at state 1. Whenever we are at state 1, there's going to be a probability, P_{1j} , that we make a transition of this kind. So out of the times that we're at state 1, there's a frequency, P_{1j} with which the next transition is into j .

So out of the overall number of transitions that happen at the trajectory, what fraction of those transitions is exactly of that kind? That fraction of transitions is the fraction of time that you find yourself at 1 times the fraction with which out of one you happen to visit next state j . So we interpreted this number as the frequency of transitions of this kind. At any given time, our chain can do transitions of different kinds, transitions of the general form from some k , I go to some l .

So we try to do some accounting. How often does a transition of each particular kind happen? And this is the frequency with which transitions of that particular kind happens. Now, what's the total frequency of transitions into state j ? Transitions into state j can happen by having a transition from 1 to j , from 2 to j , or from state m to j . So to find the total frequency with which we would observe transitions into j is going to be this particular sum.

Now, you are at state j if and only if the last transition was into state j . So the frequency with which you are at j is the frequency with which transitions into j happen. So this equation expresses exactly that statement. The probability of being at state j is the sum of the probabilities that the last transition was into state j . Or in terms of frequencies, the frequency with which you find yourself at state j is the sum of the frequencies of all the possible transition types that take you inside state j .

So that's a useful intuition to have, and we're going to see an example a little later that it gives us short cuts into analyzing Markov chains. But before we move, let's revisit the example from last time. And let us write down the balance equations for this example. So the steady-state probability that I find myself at state 1 is the probability that the previous time I was at state 1 and I made a self-transition-- So the probability that I was here last time and I made a transition of this kind, plus the probability that the last time I was here and I made a transition of that kind. So plus $\pi(2)$ times 0.2.

And similarly, for the other states, the steady-state probably that I find myself at state 2 is the probability that last time I was at state 1 and I made a transition into state 2, plus the probability that the last time I was at state 2 and I made the transition into state 1. Now, these are two equations and two unknowns, $\pi(1)$ and $\pi(2)$. But you notice that both of these equations tell you the same thing. They tell you that $0.5\pi(1)$ equals $0.2\pi(2)$.

Either of these equations tell you exactly this if you move terms around. So these two equations are not really two equations. It's just one equation. They are linearly dependent equations, and in order to solve the problem, we need the additional condition that $\pi(1) + \pi(2)$ is equal to 1. Now, we have our system of two equations, which you can solve. And once you solve it, you find that $\pi(1)$ is $2/7$ and $\pi(2)$ is $5/7$. So these are the steady state probabilities of the two different states.

If we start this chain, at some state, let's say state 1, and we let it run for a long, long time, the chain settles into steady state. What does that mean? It does not mean that the state itself enters steady state. The state will keep jumping around forever and ever. It will keep visiting both states once in a while. So the jumping never ceases. The thing that gets into steady state is the probability of finding yourself at state 1.

So the probability that you find yourself at state 1 at time one trillion is approximately $2/7$. The probability you find yourself at state 1 at time two trillions is again, approximately $2/7$. So the probability of being in that state settles into a steady value. That's what the steady-state convergence means. It's convergence of probabilities, not convergence of the process itself.

And again, the two main things that are happening in this example, and more generally, when we have a single class and no periodicity, is that the initial state does not matter. There's enough

randomness here so that no matter where you start, the randomness kind of washes out any memory of where you started. And also in this example, clearly, we do not have periodicity because we have self arcs. And this, in particular, implies that the exact time does not matter.

So now, we're going to spend the rest of our time by looking into a special class of chains that's a little easier to deal with, but still, it's an important class. So what's the moral from here? This was a simple example with two states, and we could find the steady-state probabilities by solving a simple system of two-by-two equations. If you have a chain with 100 states, it's no problem for a computer to solve a system of 100-by-100 equations. But you can certainly not do it by hand, and usually, you cannot get any closed-form formulas, so you do not necessarily get a lot of insight.

So one looks for special structures or models that maybe give you a little more insight or maybe lead you to closed-form formulas. And an interesting subclass of Markov chains in which all of these nice things do happen, is the class of birth/death processes. So what's a birth/death process? It's a Markov chain whose diagram looks basically like this. So the states of the Markov chain start from 0 and go up to some finite integer m .

What's special about this chain is that if you are at a certain state, next time you can either go up by 1, you can go down by 1, or you can stay in place. So it's like keeping track of some population at any given time. One person gets born, or one person dies, or nothing happens. Again, we're not accounting for twins here. So we're given this structure, and we are given the transition probabilities, the probabilities associated with transitions of the different types. So we use P 's for the upward transitions, Q 's for the downward transitions.

An example of a chain of this kind was the supermarket counter model that we discussed last time. That is, a customer arrives, so this increments the state by 1. Or a customer finishes service, in which case, the state gets decremented by 1, or nothing happens in which you stay in place, and so on. In the supermarket model, these P 's inside here were all taken to be equal because we assume that the arrival rate was sort of constant at each time slot. But you can generalize a little bit by assuming that these transition probabilities P_1 here, P_2 there, and so on may be different from state to state.

So in general, from state i , there's going to be a transition probability P_i that the next transition is upwards. And there's going to be a probability Q_i that the next transition is downwards. And so from that state, the probability that the next transition is downwards is going to be $Q_{(i+1)}$. So this is the structure of our chain. As I said, it's a crude model of what happens at the supermarket counter but it's also a good model for lots of types of service systems.

Again, you have a server somewhere that has a buffer. Jobs come into the buffer. So the buffer builds up. The server processes jobs, so the buffer keeps going down. And the state of the chain would be the number of jobs that you have inside your buffer.

Or you could be thinking about active phone calls out of a certain city. Each time that the phone call is placed, the number of active phone calls goes up by 1. Each time that the phone call stops happening, is terminated, then the count goes down by 1.

So it's for processes of this kind that a model with this structure is going to show up. And they do show up in many, many models. Or you can think about the number of people in a certain population that have a disease. So 1 more person gets the flu, the count goes up. 1 more person gets healed, the count goes down. And these probabilities in such an epidemic model would certainly depend on the current state.

If lots of people already have the flu, the probability that another person catches it would be pretty high. Whereas, if no one has the flu, then the probability that you get a transition where someone catches the flu, that probability would be pretty small. So the transition rates, the incidence of new people who have the disease definitely depends on how many people already have the disease. And that motivates cases where those P 's, the upward transition probabilities, depend on the state of the chain.

So how do we study this chain? You can sit down and write the system of n linear equations in the π 's. And this way, find the steady-state probabilities of this chain. But this is a little harder. It's more work than one actually needs to do. There's a very clever shortcut that applies to birth/death processes. And it's based on the frequency interpretation that we discussed a little while ago.

Let's put a line somewhere in the middle of this chain, and focus on the relation between this part and that part in more detail. So think of the chain continuing in this direction, that direction. But let's just focus on 2 adjacent states, and look at this particular cut. What is the chain going to do?

Let's say it starts here. It's going to move around. At some point, it makes a transition to the other side. And that's a transition from i to $i+1$. It stays on the other side for some time. It gets here, and eventually, it's going to make a transition to this side. Then it keeps moving and so on.

Now, there's a certain balance that must be obeyed here. The number of upward transitions through this line cannot be very different from the number of downward transitions. Because we cross this way, then next time, we'll cross that way. Then next time, we'll cross this way. We'll cross that way. So the frequency with which transitions of this kind occur has to be the same as the long-term frequency that transitions of that kind occur.

You cannot go up 100 times and go down only 50 times. If you have gone up 100 times, it means that you have gone down 99, or 100, or 101, but nothing much more different than that. So the frequency with which transitions of this kind get observed. That is, out of a large number of transitions, what fraction of transitions are of these kind? That fraction has to be the same as the fraction of transitions that happened to be of that kind.

What are these fractions? We discussed that before. The fraction of times at which transitions of this kind are observed is the fraction of time that we happen to be at that state. And out of the times that we are in that state, the fraction of transitions that happen to be upward transitions. So this is the frequency with which transitions of this kind are observed.

And with the same argument, this is the frequency with which transitions of that kind are observed. Since these two frequencies are the same, these two numbers must be the same, and

we get an equation that relates the P_i to $P_{(i+1)}$. This has a nice form because it gives us a recursion. If we knew $\pi(i)$, we could then immediately calculate $\pi(i+1)$. So it's a system of equations that's very easy to solve almost.

But how do we get started? If I knew $\pi(0)$, I could find by $\pi(1)$ and then use this recursion to find $\pi(2)$, $\pi(3)$, and so on. But we don't know $\pi(0)$. It's one more unknown. It's an unknown, and we need to actually use the extra normalization condition that the sum of the π 's is 1. And after we use that normalization condition, then we can find all of the π 's.

So you basically fix $\pi(0)$ as a symbol, solve this equation symbolically, and everything gets expressed in terms of $\pi(0)$. And then use that normalization condition to find $\pi(0)$, and you're done. Let's illustrate the details of this procedure on a particular special case. So in our special case, we're going to simplify things now by assuming that all those upward P's are the same, and all of those downward Q's are the same.

So at each point in time, if you're sitting somewhere in the middle, you have probability P of moving up and probability Q of moving down. This ρ , the ratio of P/Q is frequency of going up versus frequency of going down. If it's a service system, you can think of it as a measure of how loaded the system is. If P is equal to Q , it means that if you're at this state, you're equally likely to move left or right, so the system is kind of balanced. The state doesn't have a tendency to move in this direction or in that direction.

If ρ is bigger than 1 so that P is bigger than Q , it means that whenever I'm at some state in the middle, I'm more likely to move right rather than move left, which means that my state, of course it's random, but it has a tendency to move in that direction. And if you think of this as a number of customers in queue, it means your system has the tendency to become loaded and to build up a queue.

So ρ being bigger than 1 corresponds to a heavy load, where queues build up. ρ less than 1 corresponds to the system where queues have the tendency to drain down. Now, let's write down the equations. We have this recursion $P_{(i+1)}$ is $P_i \times \rho$ over Q_i . In our case here, the P 's and the Q 's do not depend on the particular index, so we get this relation. And this P over Q is just the load factor ρ .

Once you look at this equation, clearly you realize that by $\pi(1)$ is $\rho \times \pi(0)$. $\pi(2)$ is going to be -- So we'll do it in detail. So $\pi(1)$ is $\pi(0) \times \rho$. $\pi(2)$ is $\pi(1) \times \rho$, which is $\pi(0) \times \rho^2$. And then you continue doing this calculation. And you find that you can express every $\pi(i)$ in terms of $\pi(0)$ and you get this factor of ρ^i .

And then you use the last equation that we have -- that the sum of the probabilities has to be equal to 1. And that equation is going to tell us that the sum over all i 's from 0 to m of $\pi(0) \rho^i$ to the i is equal to 1. And therefore, $\pi(0)$ is 1 over (the sum over the ρ to the i for i going from 0 to m). So now we found $\pi(0)$, and by plugging in this expression, we have the steady-state probabilities of all of the different states.

Let's look at some special cases of this. Suppose that rho is equal to 1. If rho is equal to 1, then $\pi(i)$ is equal to $\pi(0)$. It means that all the steady-state probabilities are equal. It's means that every state is equally likely in the long run. So this is an example. It's called a symmetric random walk. It's a very popular model for modeling people who are drunk.

So you start at a state at any point in time. Either you stay in place, or you have an equal probability of going left or going right. There's no bias in either direction. You might think that in such a process, you will tend to kind of get stuck near one end or the other end. Well, it's not really clear what to expect. It turns out that in such a model, in the long run, the drunk person is equally likely to be at any one of those states.

The steady-state probability is the same for all i 's if rho is equal to 1. And so if you show up at a random time, and you ask where is my state, you will be told it's equally likely to be at any one of those places. So let's make that note. If rho equal to 1, implies that all the $\pi(i)$'s are $1/(M+1)$ -- $M+1$ because that's how many states we have in our model.

Now, let's look at a different case. Suppose that M is a huge number. So essentially, our supermarket has a very large space, a lot of space to store their customers. But suppose that the system is on the stable side. P is less than Q , which means that there's a tendency for customers to be served faster than they arrive. The drift in this chain, it tends to be in that direction. So when rho is less than 1, which is this case, and when M is going to infinity, this infinite sum is the sum of a geometric series.

And you recognize it (hopefully) -- this series is going to $1/(1-\rho)$. And because it's in the denominator, $\pi(0)$ ends up being $1-\rho$. So by taking the limit as M goes to infinity, in this case, and when rho is less than 1 so that this series is convergent, we get this formula. So we get the closed-form formula for the $\pi(i)$'s. In particular, $\pi(i)$ is $(1-\rho)(\rho^i)$.

So these $\pi(i)$'s are essentially a probability distribution. They tell us if we show up at time 1 billion and we ask, where is my state? You will be told that the state is 0. Your system is empty with probability $1-\rho$, minus or there's one customer in the system, and that happens with probability $(\rho - 1)$ times ρ . And it keeps going down this way. And it's pretty much a geometric distribution except that it has shifted so that it starts at 0 whereas the usual geometric distribution starts at 1.

So this is a mini introduction into queuing theory. This is the first and simplest model that one encounters when you start studying queuing theory. This is clearly a model of a queueing phenomenon such as the supermarket counter with the P 's corresponding to arrivals, the Q 's corresponding to departures. And this particular queuing system when M is very, very large and rho is less than 1, has a very simple and nice solution in closed form. And that's why it's very much liked.

And let me just take two seconds to draw one last picture. So this is the probability of the different i 's. It gives you a PMF. This PMF has an expected value. And the expectation, the expected number of customers in the system, is given by this formula. And this formula, which is interesting to anyone who tries to analyze a system of this kind, tells you the following.

That as long as a rho is less than 1, then the expected number of customers in the system is finite. But if rho becomes very close to 1 -- So if your load factor is something like .99, you expect to have a large number of customers in the system at any given time.

OK. All right. Have a good weekend. We'll continue next time.

MIT OpenCourseWare
<http://ocw.mit.edu>

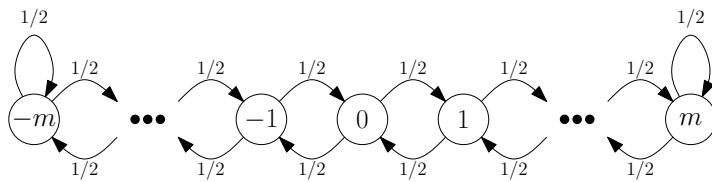
6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Problem Set 8
Due November 15, 2010

1. Oscar goes for a run each morning. When he leaves his house for his run, he is equally likely to go out either the front or back door; and similarly, when he returns, he is equally likely to go to either the front or back door. Oscar owns only five pairs of running shoes which he takes off immediately after the run at whichever door he happens to be. If there are no shoes at the door from which he leaves to go running, he runs barefooted. We are interested in determining the long-term proportion of time that he runs barefooted.
 - (a) Set the scenario up as a Markov chain, specifying the states and transition probabilities.
 - (b) Determine the long-run proportion of time Oscar runs barefooted.

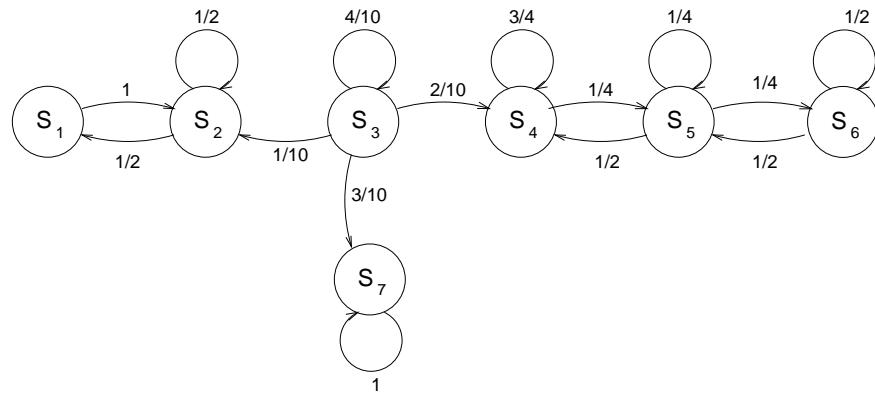
2. Consider a Markov chain X_1, X_2, \dots modeling a *symmetric simple random walk with barriers*, as shown below:



- (a) Explain why $|X_1|, |X_2|, |X_3|, \dots$ also satisfies the Markov property and draw the associated chain.
 - (b) Suppose that we also wish to keep track of the largest deviation from the origin, i.e., define the largest deviation at time t as $Y_t = \max\{|X_1|, |X_2|, \dots, |X_t|\}$. Draw a Markov chain that keeps track of the largest deviation and explain why it satisfies the Markov property.
3. As flu season is upon us, we wish to have a Markov chain that models the spread of a flu virus. Assume a population of n individuals. At the beginning of each day, each individual is either infected or susceptible (capable of contracting the flu). Suppose that each pair (i, j) , $i \neq j$, independently comes into contact with one another during the daytime with probability p . Whenever an infected individual comes into contact with a susceptible individual, he/she infects him/her. In addition, assume that overnight, any individual who has been infected for at least 24 hours will recover with probability $0 < q < 1$ and return to being susceptible, independently of everything else (i.e., assume that a newly infected individual will spend at least one restless night battling the flu).
 - (a) Suppose that there are m infected individuals at daybreak. What is the distribution of the number of new infections by day end?
 - (b) Draw a Markov chain with as few states as possible to model the spread of the flu for $n = 2$. In epidemiology, this is called an SIS (Susceptible-Infected-Susceptible) model.
 - (c) Identify all recurrent states.

Due to the nature of the flu virus, individuals almost always develop immunity after contracting the virus. Consequently, we improve our model and assume that individuals become infected at most one time. Thus, we consider individuals as either infected, susceptible, or recovered.

- (d) Draw a Markov chain to model the spread of the flu for $n = 2$. In epidemiology, this is called an SIR (Susceptible-Infected-Recovered) model.
- (e) Identify all recurrent states.
4. Consider the Markov chain below. For all parts of this problem, the process is in state 3 immediately before the first transition. Be sure to comment on any unusual results.



- (a) Find the variance for J , the number of transitions up to and including the transition on which the process leaves state 3 for the last time.
- (b) Find the expectation for K , the number of transitions up to and including the transition on which the process enters state 4 for the first time.
- (c) Find π_i for $i = 1, 2, \dots, 7$, the probability that the process is in state i after 10^{10} transitions.
- (d) Given that the process never enters state 4, find the π_i 's as defined in part (c).

G1[†]. Consider a Markov chain $\{X_k\}$ on the state space $\{1, \dots, n\}$, and suppose that whenever the state is i , a reward $g(i)$ is obtained. Let R_k be the total reward obtained over the time interval $\{0, 1, \dots, k\}$, that is, $R_k = g(X_0) + g(X_1) + \dots + g(X_k)$. For every state i , let

$$m_k(i) = E[R_k \mid X_0 = i],$$

and

$$v_k(i) = \text{var}(R_k \mid X_0 = i)$$

respectively be the conditional mean and conditional variance of R_k , conditioned on the initial state being i .

- (a) Find a recursion that, given the values of $m_k(1), \dots, m_k(n)$, allows the computation of $m_{k+1}(1), \dots, m_{k+1}(n)$.
- (b) Find a recursion that, given the values of $m_k(1), \dots, m_k(n)$ and $v_k(1), \dots, v_k(n)$, allows the computation of $v_{k+1}(1), \dots, v_{k+1}(n)$. Hint: Use the law of total variance.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

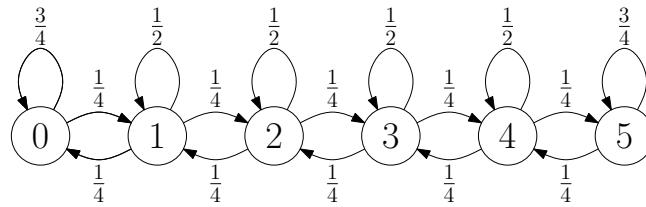
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Problem Set 8: Solutions

1. (a) We consider a Markov chain with states 0, 1, 2, 3, 4, 5, where state i indicates that there are i shoes available at the front door in the morning before Oscar leaves on his run.

Now we can determine the transition probabilities. Assuming i shoes are at the front door before Oscar sets out on his run, with probability $\frac{1}{2}$ Oscar will return to the same door from which he set out, and thus before his next run there will still be i shoes at the front door. Alternatively, with probability $\frac{1}{2}$ Oscar returns to a different door, and in this case, with equal probability there will be $\min\{i+1, 5\}$ or $\max\{i-1, 0\}$ shoes at the front door before his next run. These transition probabilities are illustrated in the following Markov chain:



- (b) When there are either 0 or 5 shoes at the front door, with probability $\frac{1}{2}$ Oscar will leave on his run from the door with 0 shoes and hence run barefooted. To find the long-term probability of Oscar running barefooted, we must find the steady-state probabilities of being in states 0 and 5, π_0 and π_5 , respectively. Note that the steady-state probabilities exist because the chain is recurrent and aperiodic.

Since this is a birth-death process, we can use the local balance equations. We have

$$\pi_0 p_{01} = \pi_1 p_{10},$$

implying that

$$\pi_1 = \pi_0$$

and similarly,

$$\pi_5 = \dots = \pi_1 = \pi_0.$$

As

$$\sum_{i=0}^5 \pi_i = 1,$$

it follows that $\pi_i = \frac{1}{6}$ for $i = 0, 1, \dots, 5$. Hence,

$$\mathbf{P}(\text{Oscar runs barefooted in the long-term}) = \frac{1}{2} (\pi_0 + \pi_5) = \frac{1}{6}.$$

2. (a) Consider any possible sequence of values $x_1, x_2, \dots, x_{t-1}, i$ for X_1, X_2, \dots, X_t , and note that

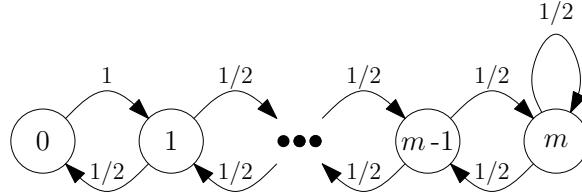
$$\mathbf{P}(|X_{t+1}| = |i| + 1 | X_t = i, X_{t-1} = x_{t-1}, \dots, X_1 = x_1) = \begin{cases} \frac{1}{2} & 0 < |i| < m \\ 1 & i = 0 \\ 0 & |i| = m \end{cases},$$

$$\mathbf{P}(|X_{t+1}| = |i| | X_t = i, X_{t-1} = x_{t-1}, \dots, X_1 = x_1) = \begin{cases} \frac{1}{2} & |i| = m \\ 0 & |i| \neq m \end{cases},$$

$$\mathbf{P}(|X_{t+1}| = |i| - 1 | X_t = i, X_{t-1} = x_{t-1}, \dots, X_1 = x_1) = \begin{cases} \frac{1}{2} & 0 < |i| \leq m \\ 0 & i = 0 \end{cases},$$

$$\mathbf{P}(|X_{t+1}| = j | X_t = i, X_{t-1} = x_{t-1}, \dots, X_1 = x_1) = 0, \quad ||i| - j| > 1.$$

As the conditional probabilities above only depend on $|i|$, where $|X_t| = |i|$, it follows that $|X_1|, |X_2|, \dots$ satisfy the Markov property. The associated Markov chain is illustrated below.



(b) Note that Y_1, Y_2, \dots is not a Markov chain for $m > 1$, because

$$\mathbf{P}(Y_{t+1} = d + 1 | Y_t = d, Y_{t-1} = d - 1) = \frac{1}{2}$$

does not equal

$$\mathbf{P}(Y_{t+1} = d + 1 | Y_t = d, Y_{t-1} = d, Y_{t-2} = d - 1) = 0,$$

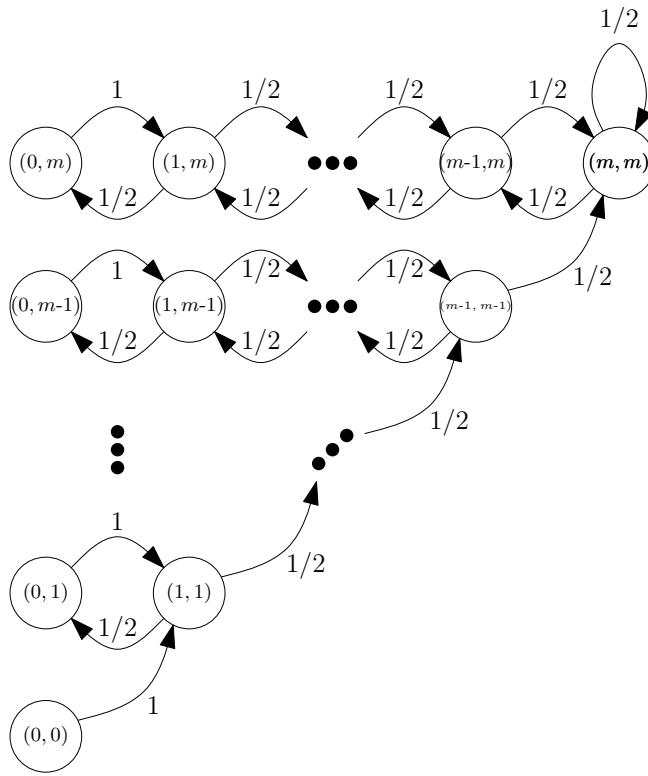
for $0 < d < m$ (the idea is that if $Y_{t-2} = d - 1, Y_{t-1} = d$, and $Y_t = d$, then $|X_t| = d - 1$, while if $Y_{t-1} = d - 1$, and $Y_t = d$, then $|X_t| = d$). If, however, we keep track of $|X_t|$ and Y_t , we do have a Markov chain, because for any possible sequence of pairs of values $(x_1, y_1), \dots, (x_{t-1}, y_{t-1}), (i_1, i_2)$ for $(|X_1|, Y_1), \dots, (|X_{t-1}|, Y_{t-1}), (|X_t|, Y_t)$,

$$\begin{aligned} \mathbf{P}((|X_{t+1}|, Y_{t+1}) = (i_1 + 1, i_2 + 1) \mid (|X_t|, Y_t) = (i_1, i_2), \dots, (|X_1|, Y_1) = (x_1, y_1)) \\ = \begin{cases} \frac{1}{2} & 0 < i_1 = i_2 < m \\ 1 & i_1 = i_2 = 0 \\ 0 & \text{otherwise} \end{cases}, \end{aligned}$$

$$\begin{aligned} \mathbf{P}((|X_{t+1}|, Y_{t+1}) = (i_1 - 1, i_2) \mid (|X_t|, Y_t) = (i_1, i_2), \dots, (|X_1|, Y_1) = (x_1, y_1)) \\ = \begin{cases} \frac{1}{2} & 0 < i_1 \leq i_2 \leq m \\ 0 & \text{otherwise} \end{cases}, \end{aligned}$$

$$\begin{aligned} \mathbf{P}((|X_{t+1}|, Y_{t+1}) = (i_1, i_2) \mid (|X_t|, Y_t) = (i_1, i_2), \dots, (|X_1|, Y_1) = (x_1, y_1)) \\ = \begin{cases} \frac{1}{2} & i_1 = i_2 = m \\ 0 & \text{otherwise} \end{cases}, \end{aligned}$$

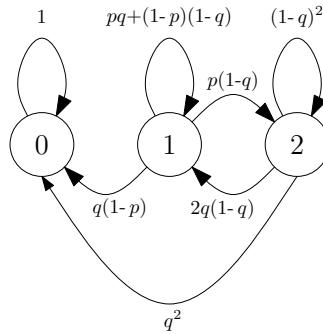
from which it is clear that the conditional probabilities only depend on (i_1, i_2) , the values of $|X_t|$ and Y_t , respectively. The corresponding Markov chain is illustrated below.



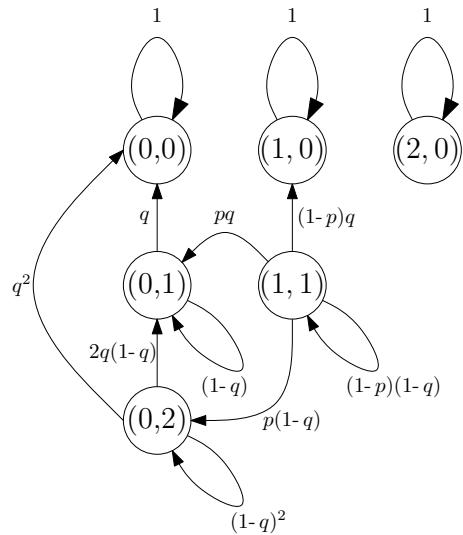
3. (a) If m out of n individuals are infected, then there must be $n - m$ susceptible individuals. Each one of these individuals will be independently infected over the course of the day with probability $\rho = 1 - (1 - p)^m$. Thus the number of new infections, I , will be a binomial random variable with parameters $n - m$ and ρ . That is,

$$p_I(k) = \binom{n-m}{k} \rho^k (1-\rho)^{n-m-k} \quad k = 0, 1, \dots, n-m.$$

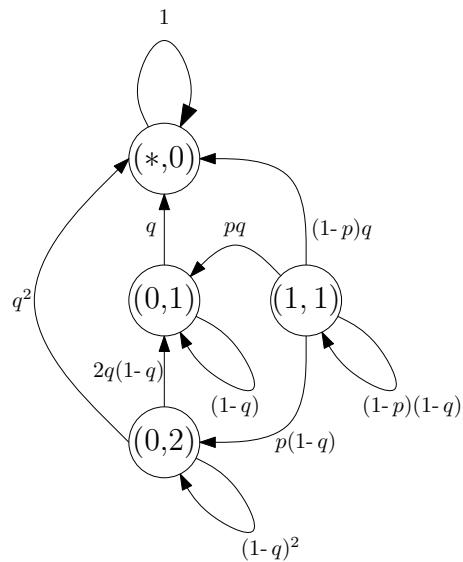
- (b) Let the state of the SIS model be the number of infected individuals. For $n = 2$, the corresponding Markov chain is illustrated below.



- (c) The only recurrent state is the state with 0 infected individuals.
 (d) Let the state of the SIR model be (S, I) , where S is the number of susceptible individuals and I is the number of infected individuals. For $n = 2$, the corresponding Markov chain is illustrated below.



If one did not wish to keep track of the breakdown of susceptible and recovered individuals when no one was infected, the three states free of infections could be consolidated into a single state as illustrated below.



- (e) Any state where the number of infected individuals equals 0 is a recurrent state. For $n = 2$, there are either one or three recurrent states, depending on the Markov chain drawn in part (d).
4. (a) The process is in state 3 immediately before the first transition. After leaving state 3 for the first time, the process cannot go back to state 3 again. Hence J , which represents the number of transitions up to and including the transition on which the process leaves state 3 for the last time is a geometric random variable with success probability equal to 0.6. The variance for J is given by:

$$\sigma_J^2 = \frac{1-p}{p^2} = \frac{10}{9}$$

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
 (Fall 2010)

- (b) There is a positive probability that we never enter state 4; i.e., $P(K < \infty) < 1$. Hence the expected value of K is ∞ .
- (c) The Markov chain has 3 different recurrent classes. The first recurrent class consists of states $\{1, 2\}$, the second recurrent class consists of state $\{7\}$ and the third recurrent class consists of states $\{4, 5, 6\}$. The probability of getting absorbed into the first recurrent class starting from the transient state 3 is,

$$\frac{1/10}{1/10 + 2/10 + 3/10} = \frac{1}{6}$$

which is the probability of transition to the first recurrent class given there is a change of state. Similarly, probability of absorption into second and third recurrent classes are $\frac{3}{6}$ and $\frac{2}{6}$ respectively.

Now, we solve the balance equations within each recurrent class, which give us the probabilities conditioned on getting absorbed from state 3 to that recurrent class. The unconditional steady-state probabilities are found by weighing the conditional steady-state probabilities by the probability of absorption to the recurrent classes.

The first recurrent class is a birth-death process. We write the following equations and solve for the conditional probabilities, denoted by p_1 and p_2 .

$$p_1 = \frac{p_2}{2}$$

$$p_1 + p_2 = 1$$

Solving these equations, we get $p_1 = \frac{1}{3}$, $p_2 = \frac{2}{3}$. For the second recurrent class, $p_7 = 1$. The third recurrent class is also a birth-death process, we can find the conditional steady-state probabilities as follows,

$$p_4 = 2p_5$$

$$p_5 = 2p_6$$

$$p_4 + p_5 + p_6 = 1$$

and thus, $p_4 = \frac{4}{7}$, $p_5 = \frac{2}{7}$, $p_6 = \frac{1}{7}$.

Using these data, the unconditional steady-state probabilities for all the states are found as follows:

$$\pi_1 = \frac{1}{3} \cdot \frac{1}{6} = \frac{1}{18}$$

$$\pi_2 = \frac{2}{3} \cdot \frac{1}{6} = \frac{1}{9}$$

$$\pi_3 = 0 \text{ (transient state)}$$

$$\pi_7 = 1 \cdot \frac{3}{6} = \frac{1}{2}$$

$$\pi_4 = \frac{4}{7} \cdot \frac{2}{6} = \frac{4}{21}$$

$$\pi_5 = \frac{2}{7} \cdot \frac{2}{6} = \frac{2}{21}$$

$$\pi_6 = \frac{1}{7} \cdot \frac{2}{6} = \frac{1}{21}$$

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
 (Fall 2010)

- (d) The given conditional event, that the process never enters state 4, changes the absorption probabilities to the recurrent classes. The probability of getting absorbed to the first recurrent class is $\frac{1}{4}$, to the second recurrent class is $\frac{3}{4}$, and to the third recurrent class is 0. Hence, the steady state probabilities are given by,

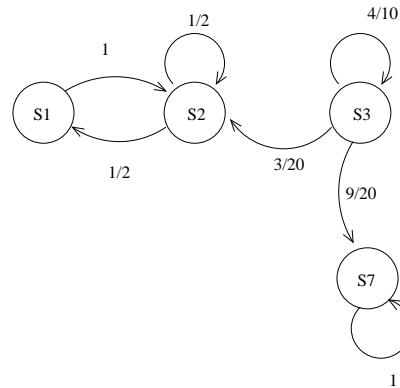
$$\pi_1 = \frac{1}{3} \cdot \frac{1}{4} = \frac{1}{12}$$

$$\pi_2 = \frac{2}{3} \cdot \frac{1}{4} = \frac{1}{6}$$

$$\pi_3 = \pi_4 = \pi_5 = \pi_6 = 0$$

$$\pi_7 = 1 \cdot \frac{3}{4} = \frac{3}{4}$$

For pedagogical purposes, let us actually draw what the new Markov chain would look like, given the event that the process never enters state 4. The resulting chain is shown below. Let us see how we came up with these transition probabilities.



We need to be careful when rescaling the new transition probabilities. First of all, it is clear that the probabilities within the recurrent classes $\{S1, S2\}$ and $\{S7\}$ don't get affected. We also note that the self loop transition probability of the transient state $S3$ doesn't get changed either.(this would be true for any other transient state)

To see that the self loop probability $p_{3,3}$ doesn't get changed, we condition on the event that we eventually enter $S2$ or $S7$. Let's call the new self loop probability, $q_{3,3}$.

Then,

$$q_{3,3} = P(X_1 = S3 \mid \text{absorbed into } 2 \text{ or } 7, X_0 = S3) = \frac{p_{3,3} * P(\text{absorbed into } 2 \text{ or } 7 \mid X_1=S3, X_0=S3)}{P(\text{absorbed into } 2 \text{ or } 7 \mid X_0=S3)}$$

$$= \frac{p_{3,3} * (a_{3,2} + a_{3,7})}{(a_{3,2} + a_{3,7})} = p_{3,3} = \frac{4}{10}$$

Now, we calculate $q_{3,7}$ and $q_{3,2}$.

$$q_{3,7} = P(X_1 = S7 \mid \text{absorbed into } 2 \text{ or } 7, X_0 = S3) = \frac{p_{3,7} * P(\text{absorbed into } 2 \text{ or } 7 \mid X_1=S7, X_0=S3)}{P(\text{absorbed into } 2 \text{ or } 7 \mid X_0=S3)}$$

$$= \frac{p_{3,7} * 1}{(a_{3,2} + a_{3,7})} = \frac{\frac{3}{10}}{\frac{1}{6} + \frac{1}{2}} = \frac{9}{20}$$

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
 (Fall 2010)

$$q_{3,2} = P(X_1 = S2 \mid \text{absorbed into 2 or 7}, X_0 = S3) = \frac{p_{3,2} * P(\text{absorbed into 2 or 7} \mid X_1 = S2, X_0 = S3)}{P(\text{absorbed into 2 or 7} \mid X_0 = S3)}$$

$$= \frac{p_{3,2} * 1}{(a_{3,2} + a_{3,7})} = \frac{\frac{1}{10}}{\frac{1}{6} + \frac{1}{2}} = \frac{3}{20}$$

Now, we can calculate the absorption probabilities of this new Markov chain.

The probability of getting absorbed into the recurrent class $\{1, 2\}$, starting from $S3$, is $\frac{\frac{3}{20}}{\frac{3}{20} + \frac{9}{20}} = \frac{1}{4}$. The probability of getting absorbed into the recurrent class $\{7\}$, starting from $S3$, is $\frac{\frac{9}{20}}{\frac{3}{20} + \frac{9}{20}} = \frac{3}{4}$. Thus, our calculated absorption probabilities match the probabilities we intuited earlier. The important thing to take away from this example is that, when doing problems of this sort, (i.e given we do/don't enter a particular set of recurrent classes), it is necessary to rescale the transition probabilities of the new chain, coming out of ALL the transient states. In other words, to find each of the new transition probabilities, we condition on the given event, that we do or do not enter particular recurrent classes.

G1[†]. a) First let the p_{ij} 's be the transition probabilities of the Markov chain.

Then

$$\begin{aligned} m_{k+1}(1) &= E[R_{k+1} \mid X_0 = 1] \\ &= E[g(X_0) + g(X_1) + \dots + g(X_{k+1}) \mid X_0 = 1] \\ &= \sum_{i=1}^n p_{1i} E[g(X_0) + g(X_1) + \dots + g(X_{k+1}) \mid X_0 = 1, X_1 = i] \\ &= \sum_{i=1}^n p_{1i} E[g(1) + g(X_1) + \dots + g(X_{k+1}) \mid X_1 = i] \\ &= g(1) + \sum_{i=1}^n p_{1i} E[g(X_1) + \dots + g(X_{k+1}) \mid X_1 = i] \\ &= g(1) + \sum_{i=1}^n p_{1i} m_k(i) \end{aligned}$$

and thus in general $m_{k+1}(c) = g(c) + \sum_{i=1}^n p_{ci} m_k(i)$ when $c \in \{1, \dots, n\}$.

Note that the third equality simply uses the total expectation theorem.

b)

$$\begin{aligned} v_{k+1}(1) &= Var[R_{k+1} \mid X_0 = 1] \\ &= Var[g(X_0) + g(X_1) + \dots + g(X_{k+1}) \mid X_0 = 1] \\ &= Var[E[g(X_0) + g(X_1) + \dots + g(X_{k+1}) \mid X_0 = 1, X_1]] + \end{aligned}$$

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
 (Fall 2010)

$$\begin{aligned}
 & E[Var[g(X_0) + g(X_1) + \dots + g(X_{k+1})|X_0 = 1, X_1]] \\
 = & Var[g(1) + E[g(X_1) + \dots + g(X_{k+1})|X_0 = 1, X_1]] + \\
 & E[Var[g(1) + g(X_1) + \dots + g(X_{k+1})|X_0 = 1, X_1]] \\
 = & Var[E[g(X_1) + \dots + g(X_{k+1})|X_0 = 1, X_1]] + E[Var[g(X_1) + \dots + g(X_{k+1})|X_0 = 1, X_1]] \\
 = & Var[E[g(X_1) + \dots + g(X_{k+1})|X_1]] + E[Var[g(X_1) + \dots + g(X_{k+1})|X_1]] \\
 = & Var[m_k(X_1)] + E[v_k(X_1)] \\
 = & E[(m_k(X_1))^2] - E[m_k(X_1)]^2 + \sum_{i=1}^n p_{1i} v_k(i) \\
 = & \sum_{i=1}^n p_{1i} m_k^2(i) - (\sum_{i=1}^n p_{1i} m_k(i))^2 + \sum_{i=1}^n p_{1i} v_k(i)
 \end{aligned}$$

so in general $v_{k+1}(c) = \sum_{i=1}^n p_{ci} m_k^2(i) - (\sum_{i=1}^n p_{ci} m_k(i))^2 + \sum_{i=1}^n p_{ci} v_k(i)$ when $c \in \{1, \dots, n\}$.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

LECTURE 17

Markov Processes – II

- **Readings:** Section 7.3

Lecture outline

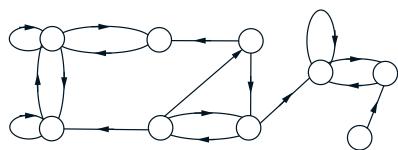
- Review
 - Steady-State behavior
 - Steady-state convergence theorem
 - Balance equations
 - Birth-death processes

Review

- Discrete state, discrete time, time-homogeneous
 - Transition probabilities p_{ij}
 - Markov property
 - $r_{ij}(n) = \mathbf{P}(X_n = j \mid X_0 = i)$
 - Key recursion:

$$r_{ij}(n) = \sum_k r_{ik}(n-1)p_{kj}$$

Warmup



$$\mathbf{P}(X_1 = 2, X_2 = 6, X_3 = 7 \mid X_0 = 1) =$$

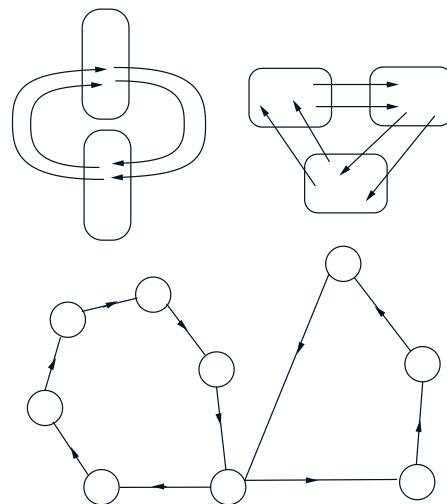
$$\mathbf{P}(X_4 = 7 \mid X_0 = 2) =$$

Recurrent and transient states

- State i is **recurrent** if:
starting from i ,
and from wherever you can go,
there is a way of returning to i
 - If not recurrent, called **transient**
 - **Recurrent class:**
collection of recurrent states that
“communicate” to each other
and to no other state

Periodic states

- The states in a recurrent class are **periodic** if they can be grouped into $d > 1$ groups so that all transitions from one group lead to the next group



Steady-State Probabilities

- Do the $r_{ij}(n)$ converge to some π_j ? (independent of the initial state i)
- Yes, if:
 - recurrent states are all in a single class, and
 - single recurrent class is not periodic
- Assuming “yes,” start from key recursion

$$r_{ij}(n) = \sum_k r_{ik}(n-1)p_{kj}$$

– take the limit as $n \rightarrow \infty$

$$\pi_j = \sum_k \pi_k p_{kj}, \quad \text{for all } j$$

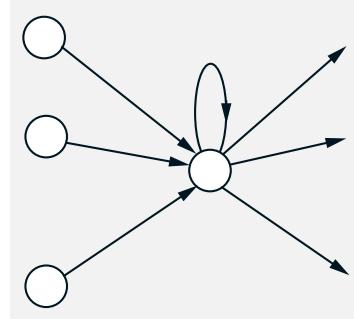
– Additional equation:

$$\sum_j \pi_j = 1$$

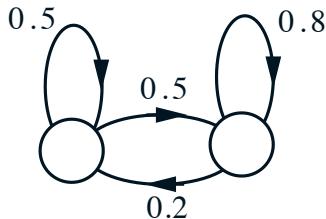
Visit frequency interpretation

$$\pi_j = \sum_k \pi_k p_{kj}$$

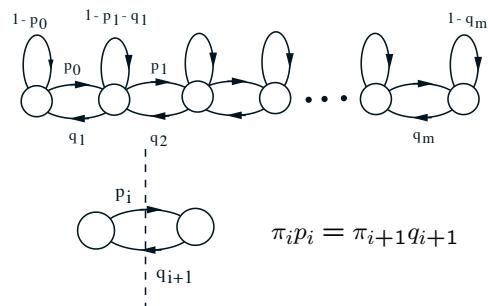
- (Long run) frequency of being in j : π_j
- Frequency of transitions $k \rightarrow j$: $\pi_k p_{kj}$
- Frequency of transitions into j : $\sum_k \pi_k p_{kj}$



Example



Birth-death processes



- Special case: $p_i = p$ and $q_i = q$ for all i
 $\rho = p/q = \text{load factor}$

$$\pi_{i+1} = \pi_i \frac{p}{q} = \pi_i \rho$$

$$\pi_i = \pi_0 \rho^i, \quad i = 0, 1, \dots, m$$

- Assume $p < q$ and $m \approx \infty$

$$\pi_0 = 1 - \rho$$

$$\mathbf{E}[X_n] = \frac{\rho}{1 - \rho} \quad (\text{in steady-state})$$

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Tutorial/Recitation 9
November 12, 2010

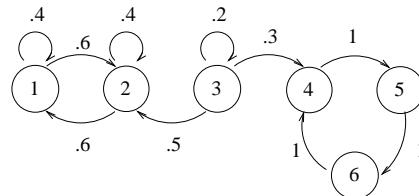
1. Problem 7.13, page 385 in textbook.

The times between successive customer arrivals at a facility are independent and identically distributed random variables with the following PMF:

$$p(k) = \begin{cases} 0.2, & k = 1 \\ 0.3, & k = 3 \\ 0.5, & k = 4 \\ 0, & \text{otherwise} \end{cases}$$

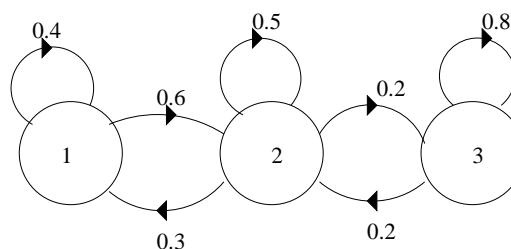
Construct a four-state Markov chain model that describes the arrival process. In this model, one of the states should correspond to the times when an arrival occurs.

2. The Markov chain shown below is in state 3 immediately before the first trial.



- (a) Indicate which states, if any, are recurrent, transient, and periodic.
 - (b) Find the probability that the process is in state 3 after n trials.
 - (c) Find the expected number of trials up to and including the trial on which the process leaves state 3.
 - (d) Find the probability that the process never enters state 1.
 - (e) Find the probability that the process is in state 4 after 10 trials.
 - (f) Given that the process is in state 4 after 10 trials, find the probability that the process was in state 4 after the first trial.
3. Problem 7.13, page 385 in textbook.

Consider the Markov chain below. Let us refer to a transition that results in a state with a higher (respectively, lower) index as a birth (respectively, death). Calculate the following quantities, assuming that when we start observing the chain, it is already in steady-state.



- (a) For each state i , the probability that the current state is i .

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Fall 2010)

- (b) The probability that the first transition we observe is a birth.
- (c) The probability that the first change of state we observe is a birth.
- (d) The conditional probability that the process was in state 2 before the first transition that we observe, given that this transition was a birth.
- (e) The conditional probability that the process was in state 2 before the first change of state that we observe, given that this change of state was a birth.
- (f) The conditional probability that the first observed transition is a birth given that it resulted in a change of state.
- (g) The conditional probability that the first observed transition leads to state 2, given that it resulted in a change of state.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.04SC Probabilistic Systems Analysis and Applied Probability

Fall 2010

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Tutorial/Recitation 9: Solutions

1. Problem 7.1, page 380 in textbook. See online solutions.
2. (a) Recurrent: 1, 2, 4, 5 , 6; Transient: 3; Periodic: 4,5,6.
(b) 0.2^n
(c) This is a geometric random variable with parameter $p = 0.5 + 0.3$. Hence, the expected number of trials up to and includ ing the trial on which the process leaves state 3 is $\mathbf{E}[X] = 1/p = 5/4$.
(d) $3/8$
(e) $\mathbf{P}(A) = 0.3 + 0.2^30.3 + 0.2^60.3 + 0.2^90.3 = 0.3024$.
(f) $0.3/\mathbf{P}(A) = 0.992$.
3. Problem 7.13, page 385 in textbook. See online solutions.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013

Transcript – Lecture 18

The following content is provided under a Creative Commons license. Your support will help MIT OpenCourseWare continue to offer high quality educational resources for free. To make a donation or view additional materials from hundreds of MIT courses, visit MIT OpenCourseWare at ocw.mit.edu.

JOHN TSITSIKLIS: So what we're going to do is to review what we have discussed last time. Then we're going to talk about the classic application of Markov chains to analyze how do you dimension a phone system. And finally, there will be two new things today. We will see how we can calculate certain interesting quantities that have to do with Markov chains. So let us start.

We've got our Markov chain and let's make the assumption that our chain is kind of nice. And by nice we mean that we've got maybe some transient states. And then we've got a single recurrent class of recurrent states. So this is a single recurrent class in the sense that from any state in that class you can get to any other state. So once you're in here you're going to circulate and keep visiting all of those states. Those states appear transient. The trajectory may move around here, but eventually one of these transitions will happen and you're going to end up in this lump.

Let's make the assumption that the single recurrent class is not periodic. These are the nicest kind of Markov chains. And they're nicest because they have the following property, the probability that you find yourself at some particular state j at the time n when that time is very large. That probability settles to a steady state value that we denote by π_j . And there are two parts in the statement. One part is that this limit exists. So the probability of state j settles to something, and furthermore that probability is not affected by i . It doesn't matter where you started, no matter where you started, the probability of state j is going to be the same in the long run. Maybe a clearer notation could be of this form. The probability of being at state j given the initial state being i is equal to π_{ij} in the limit.

Now, if I don't tell you where you started and you look at the unconditional probability of being at state i , you can average over the initial states, use the total expectation theorem and you're going to get the same answer π_i in the limit. So this tells you that the conditional probability given the initial state in the limit is the same as the unconditional probability. And that's a situation that we recognize as being one where we have independence. So what this result tells us is that X_n and X_i are approximately independent. They become independent in the limit as n goes to infinity. So that's what the steady state theorem tells us. The initial conditions don't matter, so your state at some large time n has nothing to do, is not affected by what your initial state was. Knowing the initial state doesn't tell you anything about your state at time n , therefore the states at the times-- sorry that should be a 1, or it should be a 0 -- so the state is not affected by where the process started.

So if the Markov chain is to operate for a long time and we're interested in the question where is the state, then your answer would be, I don't know, it's random. But it's going to be a particular j with this particular probability. So the steady state probabilities are interesting to us and that

raises the question of how do we compute them. The way we compute them is by solving a linear system of equations, which are called the balance equations, together with an extra equation, the normalization equation that has to be satisfied by probability, because probabilities must always add up to 1.

We talked about the interpretation of this equation last time. It's basically a conservation of probability flow in some sense. What comes in must get out. The probability of finding yourself at state j at a particular time is the total probability of the last transition taking me into state j . The last transition takes me into state j in various ways. It could be that the previous time I was at the particular state, j and i made a transition from k into j . So this number here, we interpret as the frequency with which transitions of these particular type k to j , occur. And then by adding over all k 's we consider transitions of all types that lead us inside state j . So the probability of being at the j is the sum total of the probabilities of getting into j .

What if we had multiple recurrent classes? So if we take this picture and change it to this. So here we got a secondary recurrent class. If you're here, you cannot get there. If you are here, you cannot get there. What happens in the long run? Well, in the long run, if you start from here you're going to make a transition eventually, either of this type and you would end up here, or you will make a transitional of that type and you will end up there. If you end up here, the long term statistics of your chain, that is, the probabilities of the different states, will be the steady state probabilities of this chain regarded in isolation. So you go ahead and you solve this system of equations just for this chain, and these will be your steady state probabilities. If you happened to get in here.

If, on the other hand, it happens that you went there, given that event, then what happens in the long run has to do with just this chain running by itself. So you find the steady state probabilities inside that sub chain. So you solve the linear system, the steady state equations, for this chain separately and for that chain separately. If you happen to start inside here then the steady state probabilities for this sub chain are going to apply.

Now of course this raises the question, if I start here, how do I know whether I'm going to get here or there? Well, you don't know, it's random. It may turn out that you get to here, it may turn out that you get there. So we will be interested in calculating the probabilities that eventually you end up here versus the probability that eventually you end up there. This is something that we're going to do towards the end of today's lecture.

So, as a warm up, just to see how we interpret those steady state probabilities, let us look at our familiar example. This is a 2-state Markov chain. Last time we did write down the balance equations for this chain and we found the steady state probabilities to be $2/7$ and $5/7$ respectively. So let us try to calculate some quantities.

Suppose that you start at state 1, and you want to calculate to this particular probability. So since we're assuming that we're starting at state 1, essentially here we are conditioning on the initial state being equal to 1. Now the conditional probability of two things happening is the probability that the first thing happens. But we're living in the world where we said that the initial state was

1. And then given that this thing happened, the probability that the second thing happens. But again, we're talking about conditional probabilities given that the initial state was 1.

So what is this quantity? This one is the transition probability from state 1 to state 1, so it's P_{11} . How about the second probability? So given that you started at 1 and the next time you were at 1, what's the probability that at the time 100 you are at 1? Now because of the Markov property, if I tell you that at this time you are at 1, it doesn't matter how you get there. So this part of the conditioning doesn't matter. And what we have is the 99 step transition probability from state 1 to state 1.

So the probability that you get to 1 and then 99 steps later you find yourself again at one is the probability that the first transition takes you to 1 times the probability that over the next 99 transitions starting from 1, after 99 steps you end up again at state 1. Now, 99 is possibly a big number, and so we approximate this quantity. We're using the steady state probability of state 1. And that gives us an approximation for this particular expression.

We can do the same thing to calculate something of the same kind. So you start at state 1. What's the probability that 100 steps later you are again at state 1? So that's going to be P_{11} -- not $P_{--R_{11}}$. The 100 step transition probability that starting from 1 you get to 1, and then after you get to 1 at time 100 what's the probability that the next time you find yourself at state 2? This is going to be the probability P_{12} . And approximately, since 100 is a large number, this is approximately $\pi(1)$ times P_{12} .

OK. So that's how we can use steady state probabilities to make approximations. Or you could, for example, if you continue doing examples of this kind, you could ask for what's the probability that X at time 100 is 1, and also X at time 200 is equal to 1. Then this is going to be the transition probability from 1 to 1 in 100 steps, and then over the next 100 steps from 1 you get again to 1. And this is going to be approximately $\pi(1)$ times $\pi(1)$.

So we approximate multi-step transition probabilities by the steady state probabilities when the number n that's involved in here is big. Now I said that's 99 or 100 is big. How do we know that it's big enough so that the limit has taken effect, and that our approximation is good? This has something to do with the time scale of our Markov chain, and by time scale, I mean how long does it take for the initial states to be forgotten. How long does it take for there to be enough randomness so that things sort of mix and it doesn't matter where you started?

So if you look at this chain, it takes on the average, let's say 5 tries to make a transition of this kind. It takes on the average 2 tries for a transition of that kind to take place. So every 10 time steps or so there's a little bit of randomness. Over 100 time steps there's a lot of randomness, so you expect that the initial state will have been forgotten. It doesn't matter. There's enough mixing and randomness that happens over 100 time steps. And so this approximation is good.

On the other hand, if the numbers were different, the story would have been different. Suppose that this number is 0.999 and that number is something like 0.998, so that this number becomes 0.002, and that number becomes 0.001. Suppose that the numbers were of this kind. How long does it take to forget the initial state? If I start here, there's a probability of 1 in 1,000 that next

time I'm going to be there. So on the average it's going to take me about a thousand tries just to leave that state. So, over roughly a thousand time steps my initial state really does matter.

If I tell you that you started here, you're pretty certain that, let's say over the next 100 time steps, you will still be here. So the initial state has a big effect. In this case we say that this Markov chain has a much slower time scale. It takes a much longer time to mix, it takes a much longer time for the initial state to be forgotten, and this means that we cannot do this kind of approximation if the number of steps is just 99. Here we might need n to be as large as, let's say, 10,000 or so before we can start using the approximation.

So when one uses that approximation, one needs to have some sense of how quickly does the state move around and take that into account. So there's a whole sub-field that deals with estimating or figuring out how quickly different Markov chains mix, and that's the question of when can you apply those steady state approximations.

So now let's get a little closer to the real world. We're going to talk about a famous problem that was posed, started, and solved by a Danish engineer by the name of Erlang. This is the same person whose name is given to the Erlang distribution that we saw in the context of the Poisson processes. So this was more than 100 years ago, when phones had just started existing. And he was trying to figure out what it would take to set up a phone system that how many lines should you set up for a community to be able to communicate to the outside world.

So here's the story. You've got a village, and that village has a certain population, and you want to set up phone lines. So you want to set up a number of phone lines, let's say that number is B , to the outside world. And how do you want to do that? Well, you want B to be kind of small. You don't want to set up too many wires because that's expensive. On the other hand, you want to have enough wires so that if a reasonable number of people place phone calls simultaneously, they will all get a line and they will be able to talk.

So if B is 10 and 12 people want to talk at the same time, then 2 of these people would get a busy signal, and that's not something that we like. We would like B to be large enough so that there's a substantial probability, that there's almost certainty that, under reasonable conditions, no one is going to get a busy signal.

So how do we go about modeling a situation like this? Well, to set up a model you need two pieces, one is to describe how do phone calls get initiated, and once a phone call gets started, how long does it take until the phone call is terminated? So we're going to make the simplest assumptions possible.

Let's assume that phone calls originate as a Poisson process. That is, out of that population people do not really coordinate. At completely random times, different people with decide to pick up the phone. There's no dependencies between different people, there's nothing special about different times, different times are independent. So a Poisson model is a reasonable way of modeling this situation. And it's going to be a Poisson process with some rate lambda.

Now, the rate lambda would be easy to estimate in practice. You observe what happens in that village just over a couple of days, and you figure out what's the rate at which people attempt to place phone calls. Now, about phone calls themselves, we're going to make the assumption that the duration of a phone call is a random variable that has an exponential distribution with a certain parameter μ . So $1/\mu$ is the mean duration of a phone call. So the mean duration, again, is easy to estimate. You just observe what's happening, see on the average how long these phone calls are.

Is the exponential assumption a good assumption? Well, it's means that most phone calls will be kind of short, but there's going to be a fraction of phone calls that are going to be larger, and then a very small fraction that are going to be even larger. So it sounds plausible. It's not exactly realistic, that is, phone calls that last short of 15 seconds are not that common. So either nothing happens or you have to say a few sentences and so on.

Also, back into the days when people used to connect to the internet using dial up modems, that assumption was completely destroyed, because people would dial up and then keep their phone line busy for a few hours, if the phone call was a free one. So at those times the exponential assumption for the phone call duration was completely destroyed. But leaving that detail aside, it's sort of a reasonable assumption to just get started with this problem.

All right, so now that we have those assumptions, let's try to come up with the model. And we're going to set up a Markov process model. Now the Poisson process runs in continuous time, and call durations being exponential random variables also are continuous random variables, so it seems that we are in a continuous time universe. But we have only started Markov chains for the discrete time case.

What are we going to do? We can either develop the theory of continuous time Markov chains, which is possible. But we are not going to do that in this class. Or we can discretize time and work with a discrete time model. So we're going to discretize time in the familiar way, the way we did it when we started the Poisson process. We're going to take the time axis and split it into little discrete mini slots, where every mini slot has a duration δ . So this δ is supposed to be a very small number.

So what is the state of the system? So, you look at the situation in the system at some particular time and I ask you what is going on right now, what's the information you would tell me? Well, you would tell me that right now out of these capital B lines, 10 of them are busy, or 12 of them are busy. That describes the state of the system, that tells me what's happening at this point. So we set up our states base by being the numbers from 0 to B . 0 corresponds to a state in which all the phone lines are free, no one is talking. Capital B corresponds to a case where all the phone lines are busy. And then you've got states in between.

And now let's look at the transition probabilities. Suppose that right so now we have $i-1$ lines that are busy. Or maybe, let me look here. Suppose that there's i lines that are busy. What can happen the next time? What can happen is that the new phone call gets placed, in which case my state moves up by 1, or an existing call terminates, in which case my state goes down by 1, or none of the two happens, in which case I stay at the same state. Well, it's also possible that the phone call

gets terminated and a new phone call gets placed sort of simultaneously. But when you take your time slots to be very, very small, this is going to have a negligible probability order of delta squared, so we ignore this.

So what's the probability of an upwards transition? That's the probability that the Poisson process records an arrival during a mini slot of duration delta. By the definition of the Poisson process, the probability of this happening is just lambda delta. So each one of these upwards transitions has the same probability of lambda delta. So you've got lambda deltas everywhere in this diagram.

How about, now, phone call terminations? If you had the single call that was active, so if you were here, what's the probability that the phone call terminates? So the phone call has an exponential duration with parameter mu. And we discussed before that an exponential random variable can be thought of as the first arrival time in a Poisson process. So the probability that you get this event to happen over a delta time interval is just mu times delta.

So if you have a single phone call that's happening right now, with probability mu times delta, that call is going to terminate. But suppose that we have i phone calls that are currently active. Each one of them has a probability of mu delta, of terminating, but collectively the probability that one of them terminates becomes i times mu delta. So that's because you get the mu delta contribution -- the probability of termination from each one of the different phone calls.

OK, now this is an approximate calculation, because it ignores the possibility that two phone calls terminate at the same time. Again, the way to think of why this is the correct rate, when you have i phone calls that are simultaneously running and waiting for one of them to terminate, this is like having i separate Poisson processes that are running in parallel, and you ask for the probability that one of those processes records an event. Now when you put all those processes together, it's like having a Poisson process with total rate i times mu, and so i times mu delta is the overall probability that something happens in terms of phone call terminations at those times.

So in any case, this is the transition probability for downwards transitions. Now that we've got this, we can analyze this chain. This chain has the birth death form that we discussed towards the end of last lecture. And for birth death chains, it's easy to write it out to find the steady state probabilities.

Instead of writing down the balance equations in the general form, we think in terms of a conservation of probabilities or of transitions by looking at what happens across a particular cut in this diagram. Number of transitions in the chain that cross from here to there has to be approximately equal to the number of transitions from there to here because whatever comes up must come down and then come up and so on. So the frequency with which transitions of this kind are observed has to be the same as the frequency of transitions of this kind.

What's the frequency of how often the transitions of this kind happen? And by frequency I mean quite percentage of the mini slots involve a transition of this kind? Well, for a transition of that kind to happen we need to be at states $i-1$, which happens this much of the time. And then the probability lambda delta that the transition is of this kind.

So the frequency of transitions of with which this kind of transition is observed is $\lambda \delta t$ times $\pi(i-1)$. This is the fraction of time steps at which a transition from specifically this state to specifically that state are observed. This has to be the same as the frequency with which transitions of that kind are observed, and that frequency is going to be $\mu \delta t$ times $\pi(i)$, and then we cancel the deltas, and we are left with this equation here.

So this equation expresses $\pi(i)$ in terms of $\pi(i-1)$. So if we knew $\pi(0)$ we can use that equation to determine $\pi(1)$. Once we know $\pi(1)$, we can use that equation to determine $\pi(2)$, and so on, you keep going. And the general formula that comes out of this, I will not do the algebra, it's a straightforward substitution, you find that $\pi(i)$, the steady state probability of state i is given by this expression, which involves the $\pi(0)$ from which we started.

Now what is $\pi(0)$? Well, we don't know yet, but we can find it by using the normalization equation. The sum of $\pi(i)$ has to be equal to 1. So the sum of all of those numbers has to be equal to 1. And the only way that this can happen is by setting $\pi(0)$ to be equal to that particular number. So if I tell you the value of capital B , you can set up this Markov chain, you can calculate $\pi(0)$, and then you can calculate $\pi(i)$, and so you know what fraction, you know the steady state probabilities of this chain, so you can answer the question. If I drop in at a random time, how likely is it that I'm going to find the states to be here, or the states to be there? So the steady state probabilities are probabilities, but we also interpret them as frequencies. So once I find $\pi(i)$, it also tells me what fraction of the time is the state equal to i . And you can answer that question for every possible i .

Now, why did we do this exercise? We're interested in the probability of the system is busy. So if a person, a new phone call gets placed, it just drops out of the sky. According to that Poisson process, that new phone call is going to find the system at a random state. That random state is described in steady state by the probabilities $\pi(i)$'s.

And the probability that you find the system to be busy is the probability that when you drop in the state happens to be that particular number B . So $\pi(B)$ is the probability of being busy. And this is the probability that you would like to be small in a well engineered system. So you ask the question, how should, given my λ and μ , my design question is to determine capital B the number of phone lines so that this number is small.

Could we have done, could we figure out a good value for B by doing a back of the envelope calculation? Let's suppose that λ is 30 and that μ is $1/3$. So I guess that's, let us these rates to be calls per minute. And this μ , again, is a rate per minute. Again, the units of μ are going to be calls per minute. So since our time unit is minutes, the mean duration of calls is $1/\mu$ minutes.

So a typical call, or on the average a call lasts for 3 minutes. So you get 30 calls per minute. Each call lasts for 3 minutes on the average. So on the average, if B was infinite, every call goes through. How many calls would be active on the average? So you get 30 per minute. If a call lasted exactly 1 minute, then at any time you would have 30 calls being active. Now a call lasts on the average for 3 minutes. So during each minute you generate 90 minutes of talking time.

So by thinking in terms of averages you would expect that at any time there would be about 90 calls that are active. And if 90 calls are active on the average, you could say OK, I'm going to set up my capital B to be 90. But that's not very good, because if the average number of phone calls that want to happen is if the average number is 90, sometimes you're going to have 85, sometimes you will have 95. And to be sure that the phone calls will go through you probably want to choose your capital B to be a number a little larger than 90. How much larger than 90? Well, this is a question that you can answer numerically.

So you go through the following procedure. I tried different values of capital B. For any given value of capital B, I do this numerical calculation, I find the probability that the system is busy, and then I ask what's the value of B that makes my probability of being busy to be, let's say, roughly 1 %. And if you do that calculation with the parameters that they gave you, you find that B would be something like 106.

So with the parameters they gave where you have, on the average, 90 phone calls being active, you actually need some margin to protect against the [?] fluctuation, if suddenly by chance more people want to talk, and if you want to have a good guarantee that an incoming person will have a very small probability of finding a busy system, then you will need about 106 phone lines.

So that's the calculation and the argument that the Erlang went through a long time ago. It's actually interesting that Erlang did this calculation before Markov chains were invented. So Markov's work, and the beginning of work on Markov chains, happens about 10-15 years after Erlang. So obviously he didn't call that a Markov chain. But it was something that he could study from first principles. So this is a pretty useful thing.

These probabilities that come out of that model, at least in the old days, they would all be very well tabulated in handbooks that every decent phone company engineer would sort of have with them. So this is about as practical as it gets. It's one of the sort of standard real world applications of Markov chains.

So now to close our subjects, we're going to consider a couple of new skills and see how we can calculate the few additional interesting quantities that have to do with the Markov chain. So the problem we're going to deal with here is the one I hinted that when I was talking about this picture. You start at a transient state, you're going to eventually end up here or there. We want to find the probabilities of one option of the two happening or the other happening.

So in this picture we have a class of states that's are transient. These are transient because you're going to move around those states, but there's a transition that you can make, and you go to a state from which you cannot escape afterwards. Are you going to end up here or are you going to end up there? You don't know. It's random. Let's try to calculate the probability that you end up at state 4.

Now, the probability that you end up at state 4 will depend on where you start. Because if you start here, you probably have more chances of getting to 4 because you get that chance immediately, whereas if you start here there's more chances that you're going to escape that way because it kind of takes you time to get there. It's more likely that you exit right away.

So the probability of exiting and ending up at state 4 will depend on the initial state. That's why when we talk about these absorption probability we include an index i that tells us what the initial state is. And we want to find this absorption probability, the probability that we end up here for the different initial states.

Now for some initial states this is very easy to answer. If you start at state 4, what's the probability that eventually you end up in this part of the chain? It's 1. You're certain to be there, that's where you started. If you start at state 5, what's the probability that you end up eventually at state 4? It's probability 0, there's no way to get there. Now, how about if you start at a state like state 2? If you start at state 2 then there's a few different things that can happen. Either you end up at state 4 right away and this happens with probability 0.2, or you end up at state 1, and this happens with probability 0.6. So if you end up at state 4, you are done. We are there.

If you end up at state 1, then what? Starting from state 1 there's two possibilities. Either eventually you're going to end up at state 4, or eventually you're going to end up at state 5. What's the probability of this happening? We don't know what it is, but it's what we defined to be a_1 . This is the probability -- a_1 is the probability -- that eventually you settle in state 4 given that the initial state was 1. So this probability is a_1 .

So our event of interest can happen in two ways. Either I go there directly, or I go here with probability 0.6. And given that I go there, eventually I end up at state 4, which happens with probability a_1 . So the total probability of ending up at state 4 is going to be the sum of the probabilities of the different ways that this event can happen. So our equation, in this case, is going to be, that's a_2 , is going to be 0.2 (that's the probability of going there directly) plus with probability 0.8 I end up at state 1, and then from state 1 I will end up at state 4 with probability a_1 . So this is one particular equation that we've got for what happens if we start from this state.

We can do a similar argument starting from any other state. Starting from state i the probability that eventually I end up at state 4 is, we consider the different possible scenarios of where do I go next, which is my state j , with probability P_{ij} . Next time I go to j , and given that I started at j , this is the probability that I end up at state 4.

So this equation that we have here is just an abstract version in symbols of what we wrote down for the particular case where the initial state was 2. So you write down an equation of this type for every state inside here. You'll have a separate equation for a_1 , a_2 , and a_3 . And that's going to be a system of 3 equations with 3 unknowns, the a 's inside the transient states. So you can solve that 3 by 3 system of equations. Fortunately, it turns out to have a unique solution, and so once you solve it you have found the probabilities of absorption and the probability that eventually you get absorbed at state 4.

Now, in the picture that we had here, this was a single state, and that one was a single state. How do things change if our recurrent, or trapping sets consist of multiple states? Well, it doesn't really matter that we have multiple states. All that matters is that this is one lump and once we get there we are stuck in there.

So if the picture was, let's say, like this, 0.1 and 0.2, that basically means that whenever you are in that state there's a total probability of 0.3 of ending in that lump and getting stuck inside that lump. So you would take that picture and change it and make it instead a total probability of 0.3, of ending somewhere inside that lump. And similarly, you take this lump and you view it as just one entity, and from any state you record the total probability that given that I'm here I end up in that entity. So basically, if the only thing you care is the probability that you're going to end up in this lump, you can replace that lump with a single state, view it as a single state, and calculate probabilities using this formula.

All right, so now we know where the chain is going to get to. At least we know probabilistically. We know with what probability it is going to go here, and that also tells us the probability that eventually it's going to get there. Other question, how long is it going to take until we get to either this state or that state? We can call that event absorption, meaning that the state got somewhere into a recurrent class from which it could not get out.

Okay. Let's deal with that question for the case where we have only 1 absorbing state. So here our Markov chain is a little simpler than the one in the previous slide. We've got our transient states, we've got our recurrent state, and once you get into the recurrent state you just stay there. So here we're certain that no matter where we start we're going to end up here. How long is it going to take? Well, we don't know. It's a random variable. The expected value of that random variable, let's call it μ . But how long it takes to get there certainly depends on where we start. So let's put in our notation again this index i that indicates where we started from.

And now the argument is going to be of the same type as the one we used before. We can think in terms of a tree once more, that considers all the possible options. So suppose that you start at state 1. Starting from state 1, the expected time until you end up dropping states is μ_1 . Now, starting from state 1, what are the possibilities? You make your first transition, and that first transition is going to take you either to state 2 or to state 3. It takes you to state 2 with probability 0.6, it takes you to state 3 with probability 0.4.

Starting from state 2, eventually you're going to get to state 4. How long does it take? We don't know, it's a random variable. But the expected time until this happens is μ_2 . Starting from state 2, how long does it take you to get to state 4. And similarly starting from state 3, it's going to take you on the average μ_3 time steps until you get to state 4.

So what's the expected value of the time until I end at state 4? So with probability 0.6, I'm going to end up at state 2 and from there on it's going to be expected time μ_2 , and with probability 0.4 I'm going to end up at state 3, and from there it's going to take me so much time. So this is the expected time it's going to take me after the first transition. But we also spent 1 time step for the first transition. The total time to get there is the time of the first transition, which is 1, plus the expected time starting from the next state. This expression here is the expected time starting from the next state, but we also need to account for the first transition, so we add 1. And this is going to be our μ_1 .

So once more we have a linear equation that ties together the different μ 's. And the equation starting from state 4 in this case, of course is going to be simple, starting from that state the

expected number of steps it takes you to get there for the first time is of course, 0 because you're already there. So for that state this is fine, and for all the other states you get an equation of this form.

Now we're going to have an equation for every state. It's a system of linear equations, once more we can solve them, and this gives us the expected times until our chain gets absorbed in this absorbing state. And it's nice to know that this system of equations always has a unique solution. OK so this was the expected time to absorption. For this case where we had this single absorbing state. Suppose that we have our transient states and that we have multiple recurrent classes, or multiple absorbing states.

Suppose you've got the picture like this. And we want to calculate the expected time until we get here or there. Expected time until we get to an absorbing state. What's the trick? Well, we can lump both of these states together and think of them as just one bad state, one place for which we're interested in how long it takes us to get there. So lump them as one state, and accordingly kind of merge all of those probabilities.

So starting from here, my probability that the next I end up in this lump and they get absorbed is going to be this probability plus that probability. So we would change that picture. Think of this as being just one big state. And sort of add those two probabilities together to come up with a single probability, which is the probability that starting from here next time I find myself at some absorbing state. So once you know how to deal with a situation like this, you can also find expected times to absorption for the case where you've got multiple absorbing states. You just lump all of those multiple absorbing states into a single one.

Finally, there's a kind of related quantity that's of interest. The question is almost the same as in the previous slide, except that here we do not have any absorbing states. Rather, we have a single recurrent class of states. You start at some state i . You have a special state, that is state s . And you ask the question, how long is it going to take me until I get to s for the first time? It's a single recurrent class of states. So you know that the state keeps circulating here and it keeps visiting all of the possible states. So eventually this state will be visited. How long does it take for this to happen?

Ok. So we're interested in how long it takes for this to happen, how long it takes until we get to s for the first time. And we don't care about what happens afterwards. So we might as well change this picture and remove the transitions out of s and to make them self transitions. Is the answer going to change? No. The only thing that we changed was what happens after you get to s . But what happens after you get to s doesn't matter. The question we're dealing with is how long does it take us to get to s . So essentially, it's after we do this transformation -- it's the same question as before, what's the time it takes until eventually we hit this state. And it's now in this new picture, this state is an absorbing state.

Or you can just think from first principles. Starting from the state itself, s , it takes you 0 time steps until you get to s . Starting from anywhere else, you need one transition and then after the first transition you find yourself at state j with probability P_{ij} and from then on you are going to

take expected time T_j until you get to that terminal state s . So once more these equations have a unique solution, you can solve them and find the answer.

And finally, there's a related question, which is the mean recurrence time of s . In that question you start at s , the chain will move randomly, and you ask how long is it going to take until I come back to s for the next time. So notice the difference. Here we're talking the first time after time 0, whereas here it's just the first time anywhere. So here if you start from s , T_{s^*} is not 0. You want to do at least one transition and that's how long it's going to take me until it gets back to s .

Well, how long does it take me until I get back to s ? I do my first transition, and then after my first transition I calculate the expected time from the next state how long it's going to take me until I come back to s . So all of these equations that I wrote down, they all kind of look the same. But they are different. So you can either memorize all of these equations, or instead what's better is to just to get the basic idea. That is, to calculate probabilities or expected values you use the total probability or total expectation theorem and conditional the first transition and take it from there.

So you're going to get a little bit of practice with these skills in recitation tomorrow, and of course it's in your problem set as well.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013 Transcript – Tutorial: Mean First Passage and Recurrence Times

In this problem, we are looking at a student whose performance from day to day sort of oscillates according to a Markov chain. In particular, the student can either be in state 1, which is a state of being up to date, or in state 2, which is a state of being kind of fallen behind. Now, the transition probabilities between these two states are given by the numbers here, which is 0.2 from state 1 to 2, 0.6 from 2 to 1, 0.4 from 2 back to 2, and 0.8 from 1 back to state 1.

The quantity we're interesting calculating is this notion of first passage time. Let me define what that means. Suppose we are looking at a time horizon of time 0, 1, 2, 3. And let's call the state of the Markov chain x of t . Suppose we start from the chain being in state 2 here.

Now, if we look at a particular sample path, let's say 2 and 2 again on day 1, and 2 again on day 2, and on day 3, the student enters state 1. So in this sample path, we start from time 0 and time 3 is the first time we enter state 1. And we'll say that the first passage time, namely, the first time we enter state 1 in this case, is equal to 3. More formally, we'll define t_j as the first pass the time to state 1 conditional on that we start from state j at time 0.

Now, this quantity, of course, is random. Depending on the realization, we have different numbers. And we are interested in calculating the expected value of t_2 . That is, on average, if we start from state 2 here, how long would it take for us to enter state 1?

Now to calculate this quantity, in the following recursion will be very important. The idea is we don't know exactly what t_2 is. But t_2 has to satisfy a certain recurrent equation, namely, t_2 must be equal to 1 plus summation j equal to 1 to 2 $P_{2j}t_j$.

Now let me explain what this equation means. Let's say we are at state 2. Well, we don't actually know how long it's going to take for us to enter state 1. But we do know after one step, I will be go into some other state. Let's call it state j . And from state j , it's going to take some time to enter state 1 finally. So this equation essentially says the time for us to first enter state 1 from 2 is 1-- which is the next step-- plus the expected time from that point on to enter 1. So that constitutes our [? recurrent ?] relationship.

Now, by this definition, we can see that this is simply 1 plus P_{21} times t_1 plus P_{22} times t_2 . Now, the definition of t_j says t_1 must be 0 because, by definition, if we start from state 1, we are already in state 1. So the time to reach state 1 is simply 0. So this term disappears. And we end up with 1 plus P_{22} t_2 . If we plug in a number of P_{22} -- which is 0.4 right here-- we get 1 plus 0.4 t_2 .

Now we started from t_2 and we ended up with another expression involving numbers and only one unknown, which is t_2 . Combining this together and solving for t_2 , we get t_2 equals 1 divided by 1 minus 0.4, which is 5/3. And that is the answer for the first part of the problem.

In the second part of the problem, we are asked to do something similar as before but with a slight twist. Here, I copied over the definition for t_j , which is the first time to visit state 1 starting from state j at time t equals 0. And the little t_j is this expectation.

And here we're going to define a similar quantity, which is t_1 , let's say, star, defined as the first time to visit state 1 again. So that's the recurrence part starting from state 1, 1 at t equals 0. So this is the recurrence time from state 1 back to state 1 again.

As an example, again, we look at t equals 0, 1, 2, 3, 4. And here, if we start from state 1 on time 0, we went to state 2, 2, 1, 1 again. Now here, again, time 3 will be the first time to visit state 1 after time 0. And we don't count the very first 0. And that will be our t_1 star. So t_1 star in this particular case is equal to 3. OK.

Same as before, we like to calculate the expected time to revisit state 1. Define little t_1 star expected value of t_1 star. And we'll be using the same recurrence trick through the following equation. We say that t_1 star is equal to 1 plus j from 1 to 2.

Now, since we started from state 1, this goes from 1 to state $1j$ and t_j . Again, the interpretation is we started at state 1 at time t equals 0, we went to some other state-- we call it j -- and front of state j , it goes around, and after time expected value t_j , we came back to state 1. Here, and as before, this equation works because we are working with a Markov chain whereby the time to reach some other state only depends on the current state. And that's why we're able to break down the recursion as follows.

If we write out the recursion, we get 1 plus $P_{11} t_1$ plus $P_{12} t_2$. As before, t_1 now is just the expected first passage time from state 1. And by definition, it is 0. Because if we start from state 1, it's already in state 1 and takes 0 time to get there. So again, like before, this term goes out. And we have 1 plus 0.2 times $5/3$. And this number came from the previous calculation of t_2 . And this gives us $4/3$.

So this completes the problem. And just to remind ourselves, the kind of crux of the problem is this type of recursion which expresses a certain quantity in the one incremental step followed by the expected time to reach a certain destination after that one step. And we can do so because the dynamics is modeled by a Markov chain. And hence, the time to reach a certain destination after this first step only depends on where you start again, in this case, state j .

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

LECTURE 18

Markov Processes – III

Readings: Section 7.4

Lecture outline

- Review of steady-state behavior
- Probability of blocked phone calls
- Calculating absorption probabilities
- Calculating expected time to absorption

Review

- Assume a single class of recurrent states, aperiodic; plus transient states. Then,

$$\lim_{n \rightarrow \infty} r_{ij}(n) = \pi_j$$

where π_j does not depend on the initial conditions:

$$\lim_{n \rightarrow \infty} P(X_n = j | X_0 = i) = \pi_j$$

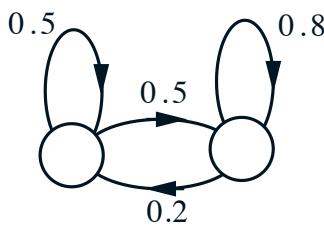
- π_1, \dots, π_m can be found as the unique solution to the balance equations

$$\pi_j = \sum_k \pi_k p_{kj}, \quad j = 1, \dots, m,$$

together with

$$\sum_j \pi_j = 1$$

Example

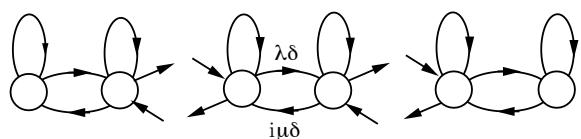


$$\pi_1 = 2/7, \pi_2 = 5/7$$

- Assume process starts at state 1.
- $P(X_1 = 1, \text{ and } X_{100} = 1) =$
- $P(X_{100} = 1 \text{ and } X_{101} = 2)$

The phone company problem

- Calls originate as a Poisson process, rate λ
 - Each call duration is exponentially distributed (parameter μ)
 - B lines available
- Discrete time intervals of (small) length δ

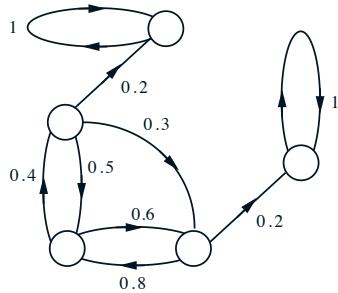


- Balance equations: $\lambda \pi_{i-1} = i \mu \pi_i$

$$\pi_i = \pi_0 \frac{\lambda^i}{\mu^i i!} \quad \pi_0 = 1 / \sum_{i=0}^B \frac{\lambda^i}{\mu^i i!}$$

Calculating absorption probabilities

- What is the probability a_i that process eventually settles in state 4, given that the initial state is i ?



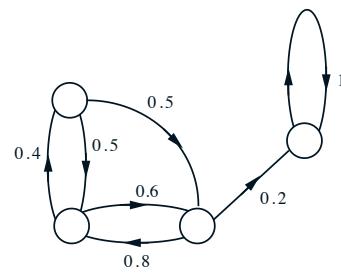
For $i = 4$, $a_i =$

For $i = 5$, $a_i =$

$$a_i = \sum_j p_{ij} a_j, \quad \text{for all other } i$$

– unique solution

Expected time to absorption



- Find expected number of transitions μ_i , until reaching the absorbing state, given that the initial state is i ?

$$\mu_i = 0 \text{ for } i =$$

$$\text{For all other } i: \mu_i = 1 + \sum_j p_{ij} \mu_j$$

– unique solution

Mean first passage and recurrence times

- Chain with one recurrent class;
fix s recurrent
- **Mean first passage time from i to s :**

$$t_i = E[\min\{n \geq 0 \text{ such that } X_n = s\} | X_0 = i]$$

- t_1, t_2, \dots, t_m are the unique solution to

$$\begin{aligned} t_s &= 0, \\ t_i &= 1 + \sum_j p_{ij} t_j, \quad \text{for all } i \neq s \end{aligned}$$

- **Mean recurrence time of s :**

$$t_s^* = E[\min\{n \geq 1 \text{ such that } X_n = s\} | X_0 = s]$$

- $t_s^* = 1 + \sum_j p_{sj} t_j$

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Fall 2010)

Recitation 19: November 16, 2010

1. Josephina is currently a 6-1 student. On each day that she is a 6-1 student, she has a probability of $1/2$ of being a course 6-1 student the next day. Otherwise, she has an equally likely chance of becoming a 6-2 student, a 6-3 student, a course 9 student or a course 15 student the next day. On any day she is a 6-3 student, she has a probability of $1/4$ of switching to course 9, a probability of $3/8$ of switching to 6-1 and a probability of $3/8$ of switching to 6-2 the next day. On any day she is a 6-2 student, she has a probability of $1/2$ of switching to course 15, a probability of $3/8$ of switching to 6-1 and a probability of $1/8$ of switching to 6-3 the next day.

In answering the questions below, assume Josephina will be a student forever. Also assume, for parts (a)-(f) that if Josephina switches to course 9 or course 15, she will stay there and will not change her course again.

- (a) What is the probability that she eventually will leave course 6?
- (b) What is the probability that she will eventually be in course 15?
- (c) What is the expected number of days until she leaves course 6?
- (d) Every time she switches into 6-1 from 6-2 or 6-3, she buys herself an ice cream cone at Tosc's. She can only afford so much ice cream, so after she's eaten 2 ice cream cones, she stops buying herself ice cream. What is the expected number of ice cream cones she buys herself before she leaves course 6?
- (e) Her friend Oscar started out just like Josephina. He is now in course 15. You don't know how long it took him to switch. What is the expected number of days it took him to switch to course 15?
- (f) Josephina decides that course 15 is not in her future. Accordingly, when she is a course 6-1 student, she stays 6-1 for another day with probability $1/2$, and otherwise she has an equally likely chance of becoming any of the other options. When she is 6-2, her probability of entering 6-1 or 6-3 are in the same proportion as before. What is the expected number of days until she is in course 9?
- (g) For this part only, assume that when Josephina is in course 9 she is equally likely to stay in course 9 or switch to course 15. Similarly, if she is in course 15, she is equally likely to stay in course 15 or switch to course 9. Calculate the probability of Josephina being in each course on any given day far into the future.
- (h) Suppose that if she is course 9 or course 15, she has probability $1/8$ of returning to 6-1, and otherwise she remains in her current course. What is the expected number of days until she is 6-1 again? (Notice that we know today she is 6-1, so if tomorrow she is still 6-1, then the number of days until she is 6-1 again is 1).

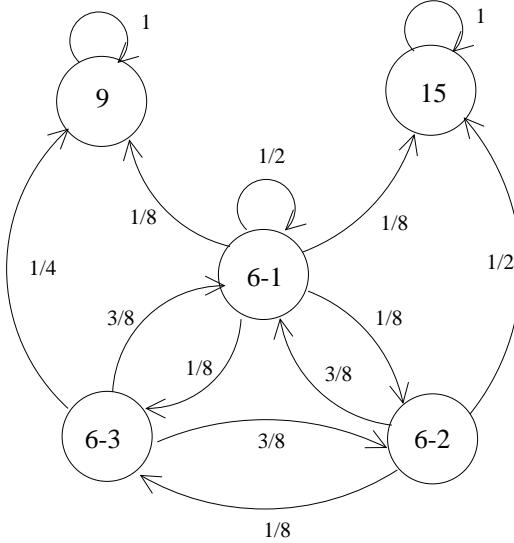
MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Recitation 19 Solutions: November 16, 2010

1. (a) The Markov chain is shown below.



By inspection, the states 6-1, 6-2, and 6-3 are all transient, since they each have paths leading to either state 9 or state 15, from which there is no return. Therefore she eventually leaves course 6 with probability $\boxed{1}$.

- (b) This is the absorption probability for the recurrent class consisting of the state course-15. Let us denote the probability of being absorbed by state 15 conditioned on being in state i as a_i . Then

$$\begin{aligned} a_{15} &= 1 \\ a_9 &= 0 \\ a_{6-1} &= \frac{1}{2}a_{6-1} + \frac{1}{8}(1) + \frac{1}{8}a_{6-2} + \frac{1}{8}(0) + \frac{1}{8}a_{6-3} \\ a_{6-2} &= \frac{1}{2}(1) + \frac{3}{8}a_{6-1} + \frac{1}{8}a_{6-3} \\ a_{6-3} &= \frac{1}{4}(0) + \frac{3}{8}a_{6-1} + \frac{3}{8}a_{6-2} \end{aligned}$$

Solving this system of equations yields

$$a_{6-1} = \frac{105}{184} \approx 0.571$$

We will keep the other a_i 's around as well - they will be useful later:

$$a_{6-2} = 0.77717$$

$$a_{6-3} = 0.50543$$

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
 (Fall 2010)

- (c) This is the expected time until absorption for the transient state $6 - 1$. Let μ_i be the expected time until absorption conditioned on being in state i . Then

$$\begin{aligned}\mu_{15} &= 0 \\ \mu_9 &= 0 \\ \mu_{6-1} &= 1 + \frac{1}{2}\mu_{6-1} + \frac{1}{8}(0) + \frac{1}{8}\mu_{6-2} + \frac{1}{8}(0) + \frac{1}{8}\mu_{6-3} \\ \mu_{6-2} &= 1 + \frac{1}{2}(0) + \frac{3}{8}\mu_{6-1} + \frac{1}{8}\mu_{6-3} \\ \mu_{6-3} &= 1 + \frac{1}{4}(0) + \frac{3}{8}\mu_{6-1} + \frac{3}{8}\mu_{6-2}\end{aligned}$$

Solving this system of equations yields

$$\mu_{6-1} = \frac{162}{46} = \frac{81}{23} \approx 3.522$$

- (d) The student buys one ice cream cone every time she goes from $6-2$ to $6-1$ or from $6-3$ to $6-1$, and buys no more than 2 ice cream cones. Let us denote $v_i(j)$ as the conditional probability that given that she is in state i , that she transitions from $6-2$ to $6-1$ or from $6-3$ to $6-1$ j additional times. Then we are interested in the expected value of the random variable N , which denotes the number of cones bought before leaving course 6, and takes on the values 0, 1, or 2. So

$$\mathbf{E}[N] = (0)v_{6-1}(0) + (1)v_{6-1}(1) + (2)(1 - v_{6-1}(0) - v_{6-1}(1))$$

We use the total probability theorem, conditioning on the next day, to yield the following set of equations:

$$\begin{aligned}v_{15}(0) &= 1 \\ v_9(0) &= 1 \\ v_{6-1}(0) &= \frac{1}{2}v_{6-1}(0) + \frac{1}{8}v_{6-2}(0) + \frac{1}{8}v_{6-3}(0) + \frac{1}{8}(1) + \frac{1}{8}(1) \\ v_{6-2}(0) &= \frac{3}{8}(0) + \frac{1}{8}v_{6-3}(0) + \frac{1}{2}(1) \\ v_{6-3}(0) &= \frac{3}{8}(0) + \frac{3}{8}v_{6-2}(0) + \frac{1}{4}(1)\end{aligned}$$

Solving this system of equations yields:

$$v_{6-1}(0) = \frac{46}{61} \approx 0.754$$

We still need to find $v_{6-1}(1)$, and we do this by again conditioning on the following day and solving the following set of equations:

$$\begin{aligned}v_{6-1}(1) &= \frac{1}{2}v_{6-1}(1) + \frac{1}{8}v_{6-2}(1) + \frac{1}{8}v_{6-3}(1) + \frac{1}{8}(0) + \frac{1}{8}(0) \\ v_{6-2}(1) &= \frac{3}{8}v_{6-1}(0) + \frac{1}{8}v_{6-3}(1) + \frac{1}{2}(0) \\ v_{6-3}(1) &= \frac{3}{8}v_{6-1}(0) + \frac{3}{8}v_{6-2}(1) + \frac{1}{4}(0)\end{aligned}$$

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
 (Fall 2010)

Notice in the second and third equations that when she transitions into state 6-1, there should be no additional transitions from 6-2 to 6-1 or from 6-3 to 6-1 after the second day in order for there to be a total of one such transition. Solving this system of equations yields:

$$v_{6-1}(1) = \frac{690}{3721} \approx 0.185$$

Finally, we can solve for the expected number of cones:

$$\begin{aligned}\mathbf{E}[N] &= (0)v_{6-1}(0) + (1)v_{6-1}(0) + (2)(1 - v_{6-1}(0) - v_{6-1}(1)) \\ &= \frac{690}{3721} + 2\left(\frac{225}{3721}\right) \\ &= \frac{1140}{3721} \approx 0.306\end{aligned}$$

- (e) We want to find the expected time to absorption conditioned on the event that the student eventually ends up in state 15, which we will call A . So

$$\begin{aligned}\mathbf{P}_{i,j|A} &= \mathbf{P}(X_{n+1} = j | X_n = i, A) \\ &= \frac{\mathbf{P}(A | X_{n+1} = j, X_n = i) \mathbf{P}(X_{n+1} = j | X_n = i)}{\mathbf{P}(A | X_n = i)} \\ &= \frac{\mathbf{P}(A | X_{n+1} = j) \mathbf{P}(X_{n+1} = j | X_n = i)}{\mathbf{P}(A | X_n = i)} \\ &= \frac{a_j \mathbf{P}_{i,j}}{a_i}\end{aligned}$$

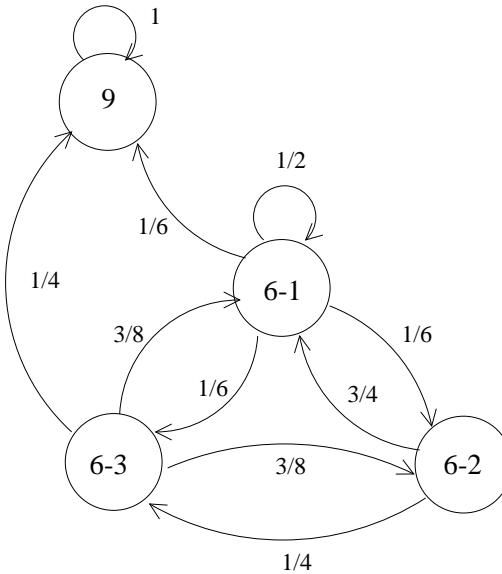
where a_k is the absorption probability of eventually ending up in state 15 conditioned on being in state k , which we found in part (b). So we may modify our chain with these new conditional probabilities and calculate the expected time to absorption on the new chain. Note that state 9 now disappears. Also, note that $\mathbf{P}_{j,j|A} = \mathbf{P}_{j,j}$, but $\mathbf{P}_{i,j|A} \neq \mathbf{P}_{i,j}$ for $i \neq j$, which means that we may not simply renormalize the transition probabilities in a uniform fashion after conditioning on this event. Let us denote the new expected time to absorption, conditioned on being in state i as $\tilde{\mu}_i$. Our system of equations now becomes

$$\begin{aligned}\tilde{\mu}_{15} &= 0 \\ \tilde{\mu}_{6-1} &= 1 + \frac{a_{6-1}}{a_{6-1}} \frac{1}{2} \tilde{\mu}_{6-1} + 0 + \frac{a_{6-2}}{a_{6-1}} \frac{1}{8} \tilde{\mu}_{6-2} + 0 + \frac{a_{6-3}}{a_{6-1}} \frac{1}{8} \tilde{\mu}_{6-3} \\ \tilde{\mu}_{6-2} &= 1 + 0 + \frac{a_{6-1}}{a_{6-2}} \frac{3}{8} \tilde{\mu}_{6-1} + \frac{a_{6-3}}{a_{6-2}} \frac{1}{8} \tilde{\mu}_{6-3} \\ \tilde{\mu}_{6-3} &= 1 + 0 + \frac{a_{6-1}}{a_{6-3}} \frac{3}{8} \tilde{\mu}_{6-1} + \frac{a_{6-2}}{a_{6-3}} \frac{3}{8} \tilde{\mu}_{6-2}\end{aligned}$$

Solving this system of equations yields

$$\tilde{\mu}_{6-1} = \frac{1763}{483} \approx 3.65$$

- (f) The new Markov chain is shown below.



This is another expected time to absorption question on the new chain. Let us define μ_k to be the expected number of days it takes the student to go from state k to state 9 in this new Markov chain:

$$\begin{aligned}\mu_{6-1} &= 1 + \frac{1}{2}\mu_{6-1} + \frac{1}{6}\mu_{6-2} + \frac{1}{6}\mu_{6-3} + \frac{1}{6}(0) \\ \mu_{6-2} &= 1 + \frac{3}{4}\mu_{6-1} + \frac{1}{4}\mu_{6-3} \\ \mu_{6-3} &= 1 + \frac{3}{8}\mu_{6-1} + \frac{3}{8}\mu_{6-2} + \frac{1}{4}(0)\end{aligned}$$

Solving this system of equations yields:

$$\mu_{6-1} = \frac{86}{13} \approx 6.615$$

- (g) States 6-1, 6-2 and 6-3 are now transient. States 9 and 15 form a recurrent class. By symmetry, 9 and 15 have the same steady state probability of 1/2.

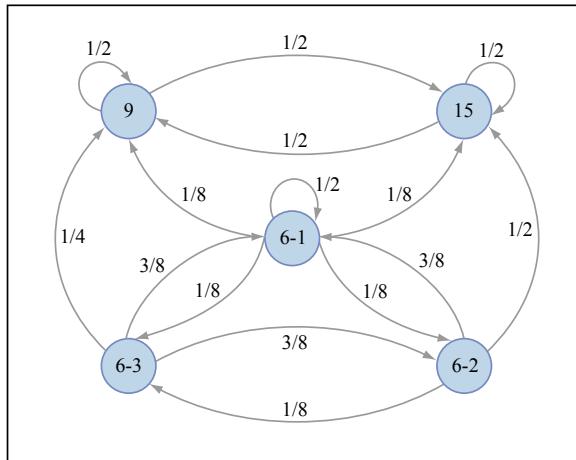


Image by MIT OpenCourseWare.

States 6-1, 6-2 and 6-3 are now transient. States 9 and 15 form a recurrent class. By symmetry, 9 and 15 have the same steady state probability of 1/2.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
 (Fall 2010)

- (h) The corresponding Markov chain is the same as the one in part (a) except $p_{9,6-1} = \frac{1}{8}, p_{9,9} = \frac{7}{8}, p_{15,6-1} = \frac{1}{8}, p_{15,15} = \frac{7}{8}$ instead of $p_{9,9} = 1, p_{15,15} = 1$.

We can consider state 6-1 as an absorbing state. Let μ_k be the expected number of transitions until absorption if we start at state k

$$\begin{aligned}\mu_9 &= \frac{1}{8} + \frac{7}{8}(1 + \mu_9) \Rightarrow \mu_9 = 8 \\ \mu_{15} &= \frac{1}{8} + \frac{7}{8}(1 + \mu_{15}) \Rightarrow \mu_{15} = 8 \\ \mu_{6-3} &= \frac{3}{8} + \frac{3}{8}(1 + \mu_{6-2}) + \frac{1}{4}(1 + \mu_9) \\ \mu_{6-2} &= \frac{3}{8} + \frac{1}{8}(1 + \mu_{6-3}) + \frac{1}{2}(1 + \mu_{15}) \\ &\Rightarrow \mu_{6-2} = \frac{344}{61}, \mu_{6-3} = \frac{312}{61}\end{aligned}$$

Let R be the number of days until she is 6-1 again. We find $E[R]$ by using the total expectation theorem, conditioned on what happens on the first transition.

$$\begin{aligned}\mathbf{E}[R] &= \mathbf{E}[\mathbf{E}[R|X_2]] \\ &= \frac{1}{2}(1) + \frac{1}{8}(1 + \mu_9) + \frac{1}{8}(1 + \mu_{15}) + \frac{1}{8}(1 + \mu_{6-2}) + \frac{1}{8}(1 + \mu_{6-3}) \\ &= \frac{265}{61}\end{aligned}$$

Notice that this chain consists of a single recurrent aperiodic class. Another approach to solving this problem uses the steady state probabilities of this chain, which are $\pi_{6-1} = \frac{61}{265}, \pi_{6-2} = \frac{11}{265}, \pi_{6-3} = \frac{9}{265}, \pi_9 = \frac{79}{265}, \pi_{15} = \frac{105}{265}$. The expected frequency of visits to 6-1 is π_{6-1} , so the expected number of days between visits to 6-1 is $\frac{1}{\pi_{6-1}}$.¹ Since she is currently 6-1, the expected number of days until she is 6-1 again is $\frac{1}{\pi_{6-1}} = \frac{265}{61}$.

¹See problem 7.34 on page 399 of the text for a more detailed explanation of this correspondence between mean recurrence times and steady-state probabilities.

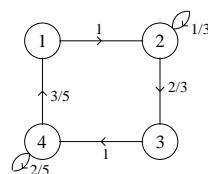
MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Tutorial 10
November 18/19, 2010

1. Define X as the height in meters of a randomly selected Canadian, where the selection probability is equal for each Canadian, and denote $\mathbf{E}[X]$ by h . Bo is interested in estimating h . Because he is sure that no Canadian is taller than 3 meters, Bo decides to use 1.5 meters as a conservative (large) value for the standard deviation of X . To estimate h , Bo averages the heights of n Canadians that he selects at random; he denotes this quantity by H .
 - (a) In terms of h and Bo's 1.5 meter bound for the standard deviation of X , determine the expected value and standard deviation for H .
 - (b) Help Bo by calculating a minimum value of n (with $n > 0$) such that the standard deviation of Bo's estimator, H , will be less than 0.01 meters.
 - (c) Bo would like to be 99% sure that his estimate is within 5 centimeters of the true average height of Canadians. Using the Chebyshev inequality, calculate the minimum value of n that will make Bo happy.
 - (d) If we agree that no Canadians are taller than three meters, why is it correct to use 1.5 meters as an upper bound on the standard deviation for X , the height of any Canadian selected at random?
2. On any given week while taking 6.041, a student can be either up-to-date on learning the material, or she may have fallen behind. If she is up-to-date in a given week, the probability that she will be up-to-date (or behind) in the next week is 0.8 (or 0.2, respectively). If she is behind in the given week, the probability that she will be up-to-date (or behind) in the next week is 0.6 (or 0.4, respectively). We assume that these probabilities do not depend on whether she was up-to-date or behind in previous weeks, so we can model the situation as a 2-state Markov chain where State 1 is the case when the student is up-to-date and State 2 is the case when the student is behind.
 - (a) Calculate the mean first passage time to State 1, starting from State 2.
 - (b) Calculate the mean recurrence time to State 1.
3. Consider the following Markov chain:



The steady-state probabilities for this process are:

$$\pi_1 = \frac{6}{31} \quad \pi_2 = \frac{9}{31} \quad \pi_3 = \frac{6}{31} \quad \pi_4 = \frac{10}{31}$$

Assume the process is in state 1 just before the first transition.

- (a) Determine the expected value and variance of K , the number of transitions up to and including the next transition on which the process returns to state 1.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Fall 2010)

- (b) What is the probability that the state of the system resulting from transition 1000 is neither the same as the state resulting from transition 999 nor the same as the state resulting from transition 1001?

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Tutorial 10 Solutions
November 18/19, 2010

1. Note that n is deterministic and H is a random variable.

(a) Use X_1, X_2, \dots to denote the (random) measured heights.

$$\begin{aligned} H &= \frac{X_1 + X_2 + \cdots + X_n}{n} \\ \mathbf{E}[H] &= \frac{\mathbf{E}[X_1 + X_2 + \cdots + X_n]}{n} = \frac{n\mathbf{E}[X]}{n} = h \\ \sigma_H &= \sqrt{\text{var}(H)} = \sqrt{\frac{n \text{var}(X)}{n^2}} \quad (\text{var of sum of independent r.v.s is sum of vars}) \\ &= \frac{1.5}{\sqrt{n}} \end{aligned}$$

(b) We solve $\frac{1.5}{\sqrt{n}} < 0.01$ for n to obtain $n > 22500$.

(c) Apply the Chebyshev inequality to H with $\mathbf{E}[H]$ and $\text{var}(H)$ from part (a):

$$\begin{aligned} \mathbf{P}(|H - h| \geq t) &\leq \left(\frac{\sigma_H}{t}\right)^2 \\ \mathbf{P}(|H - h| < t) &\geq 1 - \left(\frac{\sigma_H}{t}\right)^2 \end{aligned}$$

To be “99% sure” we require the latter probability to be at least 0.99. Thus we solve

$$1 - \left(\frac{\sigma_H}{t}\right)^2 \geq 0.99$$

with $t = 0.05$ and $\sigma_H = \frac{1.5}{\sqrt{n}}$ to obtain

$$n \geq \left(\frac{1.5}{0.05}\right)^2 \frac{1}{0.01} = 90000.$$

(d) Intuitively, the variance of a random variable X that takes values in the range $[0, b]$ is maximum when X takes the value 0 with probability 0.5 and the value b with probability 0.5, in which case the variance of X is $b^2/4$ and its standard deviation is $b/2$.

More formally, since $\mathbf{E}[(X - c)^2]$ is minimized when $c = \mathbf{E}[X]$, we have for any random variable X taking values in $[0, b]$,

$$\begin{aligned} \text{var}(X) &\leq \mathbf{E}[(X - \frac{b}{2})^2] \\ &= \mathbf{E}[X^2] - b\mathbf{E}[X] + \frac{b^2}{4} \\ &= \mathbf{E}[X(X - b)] + \frac{b^2}{4} \\ &\leq 0 + \frac{b^2}{4}, \end{aligned}$$

since $0 \leq X \leq b \Rightarrow X(X - b) \leq 0$. Thus $\sigma_X \leq b/2$.

In our example, we have $b = 3$, so $\sigma_X \leq 3/2$.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
 (Fall 2010)

2. (a) Setting $s = 1$, we get $t_1 = 0$ and

$$\begin{aligned} t_2 &= 1 + \sum_{j=1}^m p_{ij}t_j \quad \forall i \neq s, \\ &= 1 + p_{22}t_2 \\ &\Rightarrow t_2 = 5/3. \end{aligned}$$

(b)

$$\begin{aligned} t_s^* &= 1 + \sum_{j=1}^m p_{sj}t_j \\ t_1^* &= 1 + p_{12}t_2 = 4/3. \end{aligned}$$

3. (a) $K = 2 + X_1 + X_2$, where X_1 and X_2 are independent exponential random variables with parameters $2/3$ and $3/5$.

$$\begin{aligned} E[K] &= 2 + 1/p_1 + 1/p_2 \\ &= 31/6. \\ \text{var}(K) &= \frac{1-p_1}{p_1^2} + \frac{1-p_2}{p_2^2} \\ &= 67/36. \end{aligned}$$

(b)

$$\begin{aligned} \mathbf{P}(A) &= \mathbf{P}(X_{999} \neq X_{1000} \neq X_{1001}) \\ &= \sum_{i=1}^4 \mathbf{P}(A|X_{999} = i)\pi_i \\ &= 2/3\pi_1 + 2/3\pi_2 + 3/5\pi_3 + 3/5\pi_4 \\ &= 30/93 + 48/155 \approx 0.6323. \end{aligned}$$

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013 Transcript – Recitation: Convergence in Probability and in the Mean Part 1

In this exercise, we'll be working with the notion of convergence in probability, as well as some other notion of converge of random variables that we'll introduce later. First type of random variable is x_n , where x_n has probability $1 - \frac{1}{n}$ to be as 0 and probability of $\frac{1}{n}$ to be a 1. And graphically, we see that we have a pretty big mess. $1 - \frac{1}{n}$ at location 0, and a tiny bit somewhere here, only $\frac{1}{n}$. So this will be the PMF for x .

On the other hand, we have the sequence of random variables, y_n . Fairly similar to x_n with a slight tweak. The similar part says it also has a very high probability of being at 0, mass $\frac{1}{n}$. But on the off chance that y_n is not at 0, it has a pretty big value n . So it has probability $\frac{1}{n}$ of somewhere out there. So to contrast the two graphs, we see at 0, they have the same amount of mass, $\frac{1}{n}$, but for y , it's all the way out there that has a small mass $\frac{1}{n}$. So this will be our PMF for y .

And for the remainder of the problem, we'll be looking at the regime where the number n tends to infinity, and study what will happen to these two sequences of random variables. In part A, we're to compute the expected value and variance for both x_n and y_n . Let's get started.

The expected value of x_n is given by the probability-- it's at one, which is $\frac{1}{n}$ times 1 plus the probability being at 0, $\frac{1}{n}$ times value 0. And that gives us $\frac{1}{n}$. To calculate the variance of x_n , see that variance is simply the expected value of x_n minus expected value of x_n , which in this case is $\frac{1}{n}$ from the previous calculation we have here.

We take the square of this value and compute the whole expectation, and this gives us $\frac{1}{n} - \frac{1}{n^2}$ plus the remainder probability $1 - \frac{1}{n}$ of x being at 0, so $0 - \frac{1}{n^2}$. And if we carry out the calculations here, we'll get $\frac{n-1}{n^2}$.

Now, let's turn to y_n . The expected value of y_n is equal to probability of being at 0, 0 plus the probability of being at n and times the value n . And this gives us 1. The variance of y_n . We do the same thing as before, we have $1 - \frac{1}{n}$ probability of being at 0, multiplied $0 - \frac{1}{n}$ squared, where 1 is the expected value of y . And with probability $\frac{1}{n}$, out there, equal to n , and this is multiplied by $n - \frac{1}{n}$ squared. And this gives us $\frac{n-1}{n}$.

Already, we can see that while the expected value for x was $\frac{1}{n}$, the expected value for y is sitting right at 1. It does not decrease as it increases. And also, while the variance for x is $\frac{n-1}{n^2}$, the variance for y is much bigger. It is actually increasing to infinity as n goes infinity. So these intuitions will be helpful for the remainder of the problem.

In part B, we're asked to use Chebyshev's Inequality and see whether x_n or y_n converges to any number in probability. Let's first recall what the inequality is about. It says that if we have random variable x , in our case, x_n , then the probability of x_n minus the expected value of x_n , in our case, $\frac{1}{n}$, that this deviation, the absolute value of this difference is greater than ϵ is bounded above by the variance of x_n divided by the value of ϵ squared.

Well, in our case, we know the variance is n minus 1 over n squared, hence this whole term is this term divided by epsilon squared. Now, we know that as n gets really big, the probability of x_n being at 0 is very big. It's 1 minus 1 over n . So a safe bet to guess is that if x_n work to converge anywhere on the real line, it might just converge to the point 0. And let's see if that is true.

Now, to show that x_n converges to 0 in probability, formally we need to show that for every fixed epsilon greater than 0, the probability that x_n minus 0 greater than epsilon has to be 0, and the limit has n going to infinity. And hopefully, the inequalities above will help us achieve this goal. And let's see how that is done.

I would like to have an estimate, in fact, an upper bound of the probability x_n absolute value greater or equal to epsilon. And now, we're going to do some massaging to this equation so that it looks like what we know before, which is right here. Now, we see that this equation is in fact, less than probability x_n minus 1 over n greater or equal to epsilon plus 1 over n .

Now, I will justify this inequality in one second. But suppose that you believe me for this inequality, we can simply plug-in the value right here, namely substituting epsilon plus 1 over n , in the place of epsilon right here and use the Chebyshev Inequality we did earlier to arrive at the following inequality, which is n minus 1 over n squared times, instead of epsilon, now we have epsilon plus 1 over n squared.

Now, if we take n to infinity in this equation, see what happens. Well, this term here converges to 0 because n squared is much bigger than n minus 1. And this term here converges to number 1 over epsilon squared. So it becomes 0 times 1 over epsilon squared, hence the whole term converges to 0. And this proves that indeed, the limit of the term here as n going to infinity is equal to 0, and that implies x_n converges to 0 in probability.

Now, there is the one thing I did not justify in the process, which is why is probability of absolute value x_n greater than epsilon less than the term right here? So let's take a look. Well, the easiest way to see this is to see what ranges of x_n are we talking about in each case.

Well, in the first case, we're looking at interval around 0 plus minus epsilon and x_n can lie anywhere here. While in the second case, right here, we can see that the set of range values for x_n is precisely this interval here, which was the same as before, but now, we actually have less on this side, where the starting point and the interval on the right is epsilon plus 2 over n . And therefore, the right hand style captures strictly less values of x_n than the left hand side, hence the inequality is true.

Now, we wonder if we can use the same trick, Chebyshev Inequality, to derive the result for y_n as well. Let's take a look. The probability of y_n minus its mean, 1, greater or equal to epsilon. From the Chebyshev Inequality, we have variance of y_n divided by epsilon squared.

Now, there is a problem. The variance of y_n is very big. In fact, it is equal to n minus 1. And we calculated in part A, divided by epsilon squared. And this quantity here diverges as n going to

infinity to infinity itself. So in this case, the Chebyshev Inequality does not tell us much information of whether y_n converges or not.

Now, going to part C, the question is although we don't know anything about y_n from just the Chebyshev Inequality, does y_n converge to anything at all? Well, it turns out it does. In fact, we don't have to go through anything more complicated than distribution y_n itself.

So from the distribution y_n , we know that absolute value of y_n greater or equal to epsilon is equal to 1 over n whenever epsilon is less than n . And this is true because we know y_n has a lot of mass at 0 and a tiny bit a mass at value 1 over n at location n .

So if we draw the cutoff here at epsilon, then the probability of y_n landing to the right of epsilon is simply equal to 1 over n . And this tells us, if we take the limit of n going to infinity and measure the probability that y_n -- just to write it clearly-- deviates from 0 by more than epsilon, this is equal to the limit as n going to infinity of 1 over n . And that is equal to 0. From this calculation, we know that y_n does converge to 0 in probability as n going to infinity.

For part D, we'd like to know whether the convergence in probability implies the convergence in expectation. That is, if we have a sequence of random variables, let's call it z_n , that converges to number c in probability as n going to infinity, does it also imply that the limit as n going to infinity of the expected value of z_n also converges to c . Is that true?

Well, intuitively it is true, because in the limit, z_n almost looks like it concentrates on c solely, hence we might expect that the expected value is also going to c itself. Well, unfortunately, that is not quite true. In fact, we have the proof right here by looking at y_n . For y_n , we know that the expected value of y_n is equal to 1 for all n . It does not matter how big n gets. But we also know from part C that y_n does converge to 0 in probability.

And this means somehow, y_n can get very close to 0, yet its expected value still stays one away. And the reason again, we go back to the way y_n was constructed. Now, as n goes to infinity, the probability of y_n being at 0, 1 minus 1 over n , approaches 1.

So it's very likely that y_n is having a value 0, but whenever on the off chance that y_n takes a value other than 0, it's a huge number. It is n , even though it has a small probability of 1 over n . Adding these two factors together, it tells us the expected value of y_n always stays at 1.

And however, in probability, it's very likely that y is around 0. So this example tells us converges in probability is not that strong. That tells us something about the random variables but it does not tell us whether the mean value of the random variables converge to the same number.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013 Transcript – Recitation: Convergence in Probability and in the Mean Part 2

For part E and F of the problem, we'll be introducing a new notion of convergence, so-called the convergence E mean squared sense. We say that x_n converges to a number c in mean squared, if as we take and go to infinity, the expected value of x_n minus c squared goes to 0. To get a sense of what this looks like, let's say we let c equal to the expected value of x_n , and let's say the expected value of x_n is always the same.

So the sequence of random variables has the same mean. Well, if that is true, then mean square convergence simply says the limit of the variance of x_n is 0. So as you can imagine, somehow as x_n becomes big, the variance of x_n is very small, so x_n is basically highly concentrated around c . And by this I mean, the density function for x_n . So that's the notion of convergence we'll be working with.

Our first task here is to show that the mean square convergence is in some sense stronger than the convergence in probability that we have been working with from part A to part D. That is, if I know that x_n converged to some number c in mean squared, then this must imply that x_n converges to c in probability. And now, we'll go show that for part E.

Well, let's start with a definition of convergence in probability. We want to show that for a fixed constant ϵ the probability that x_n minus c , greater than ϵ , essentially goes to 0 as n goes to infinity. To do so, we look at the value of this term.

Well, the probability of absolute value x_n minus c greater than ϵ is equal to the case if we were to square both sides of the inequality. So that is equal to the probability that x_n minus c squared greater than ϵ^2 . We can do this because both sides are positive, hence this goes through.

Now, to bound this equality, we'll invoke the Markov's Inequality, which it says this probability of x_n , some random variable greater than ϵ^2 , is no more than is less equal to the expected value of the random variable. In this case, the expected value of x_n minus c squared divided by the threshold that we're trying to cross. So that is Markov's Inequality.

Now, since we know x_n converges to c in mean squared, and by definition, mean square we know this precise expectation right here goes to 0. And therefore, the whole expression goes to 0 as n goes to infinity. Because the denominator here is a constant and the top, the numerator here, goes to 0. So now we have it. We know that the probability of x_n minus c absolute value greater than ϵ goes to 0 as n goes to infinity, for all fixed value of ϵ s and this is the definition of convergence in probability.

Now that we know if x_n converges to c mean squared, it implies that x_n converges to c in probability. One might wonder whether the reverse is true. Namely, if we know something converges in probability to a constant, does the same sequence of random variables converge to

the same constant in mean squared? It turns out that is not quite the case. The notion of probability converges in probability is not as strong as a notion of convergence in mean squared.

Again, to look for a counter example, we do not have to go further than the y_n 's we have been working with. So here we know that y_n converges to 0 in probability. But it turns out it does not converge to 0 in the mean squared. And to see why this is the case, we can take the expected value of y_n minus 0 squared, and see how that goes.

Well, the value of this can be computed easily, which is simply 0, if y_n is equal to 0, with probability 1 minus n plus n squared when y_n takes a value of n , and this happens with probability 1 over n . The whole expression evaluates to n , which blows up to infinity as n going to infinity. As a result, the limit n going to infinity of E of y_n minus 0 squared is infinity and is not equal to 0. And there we have it, even though y_n converges to 0 in probability, because the variance of y_n , in some sense, is too big, it does not converge in a mean squared sense.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013 Transcript – Recitation: Convergence in Probability Example

In this problem, we're given a random variable X which has a uniform distribution in the interval negative 1 to 1. In other words, if we were to draw out the PDF of X , we see that in the interval negative 1 to 1, it has value 1/2. Now we're given a sequence random variables X_1, X_2 , and so on, where each X_i has the same distribution as X and different X_i 's are independent.

For part a, we would like to know if the sequence X_i converges to some number-- let's call it c -- in probability as i goes to infinity-- whether this is true. Let's first recall the definition of convergence in probability. If this does happen, then by definition, we'll have that for every epsilon greater than 0, the probability $|X_i - c| \geq \epsilon$ is less than or equal to epsilon, this quantity will go to 0 in the limit of i going to infinity. In other words, with very high probability, we will find X_i to be very concentrated around the number c if this were to be the PDF of X_i .

Now, can this be true? Well, we know that each X_i is simply a uniform distribution over negative 1 to 1. It doesn't really change as we increase i . So intuitively, the concentration around any number c is not going to happen. So we should not expect a convergence in probability in this sense.

For part b, we would like to know whether the sequence Y_i , defined as X_i divided by i , converges to anything in probability. Well, by just looking at the shape of Y_i , we know that since the absolute value of X_i is less than 1, then we expect the absolute value of Y_i is less than 1/ i . So eventually, Y_i gets very close to 0 as i goes to infinity. So it's safe to bet that maybe Y_i will converge to 0 in probability.

Let's see if this is indeed the case. The probability of $|Y_i| > \epsilon$ is equal to the probability of $|X_i| / i > \epsilon$. Now, previously we know that the absolute value of X_i is at most 1 by the definition of Y_i . And hence the probability right here is upper bounded by the probability of $|X_i| > i\epsilon$.

Notice in this expression, there is nothing random. i is simply a number. Hence this is either 1 if i is less than or equal to $1/\epsilon$, or 0 if i is greater than $1/\epsilon$.

Now, this tells us, as long as i is great enough-- it's big enough compared to ϵ -- we know that this quantity here is [INAUDIBLE] 0. And that tells us in the limit of i goes to infinity probability of Y_i deviating from 0 by more than ϵ goes to 0. And that shows that indeed, Y_i converges to 0 in probability because the expression right here, this limit, holds for all ϵ .

Now, in the last part of the problem, we are looking at a sequence Z_i defined by X_i raised to the i -th power. Again, since we know X_i is some number between negative 1 and 1, this number raised to the i -th power is likely to be very small. And likely to be small in the sense that it will have absolute value close to 0. So a safe guess will be the sequence Z_i converges to 0 as well as i goes to infinity.

How do we prove this formally? We'll start again with a probability that Z_i stays away from 0 by more than epsilon and see how that evolves. And this is equal to the probability that X_i raised to the i -th power greater equal to epsilon. Or again, we can write this by taking out the absolute value that X_i is less equal to negative epsilon raised to the 1 over i -th power or X_i greater equal to epsilon 1 over i -th power.

So here, we'll divide into two cases, depending on the value of epsilon. In the first case, epsilon is greater than 1. Well, if that's the case, then we know epsilon raised to some positive power is still greater than 1. But again, X_i cannot have any positive density be on the interval negative 1 or 1. And hence we know the probability above, which is X_i less than some number smaller than negative 1 or greater than some number bigger than 1 is 0. So that case is handled.

Now let's look at a case where epsilon is less than 1, greater than 0. So in this case, epsilon to the $1/i$ will be less than 1. And it's not that difficult to check that since X_i has uniform density between negative 1 and 1 of magnitude $1/2$, then the probability here was simply 2 times $1/2$ times the distance between epsilon to the 1 over i -th power and 1.

So in order to prove this quantity converge to 0, we simply have to justify why does epsilon to the $1/i$ converge to 1 as i goes to infinity. For that, we'll recall the properties of exponential functions. In particular, if a is a positive number and x is its exponent, if we were to take the limit as x goes to 0 and look at the value of a to the power of x , we see that this goes to 1.

So in this case, we'll let a be equal to epsilon and x be equal to $1/i$. As we can see that as i goes to infinity, the value of x , which is $1/i$, does go to 0. And therefore, in the limit i going to infinity, the value of epsilon to the 1 over i -th power goes to 1.

And that shows if we plug this limit into the expression right here that indeed, the term right here goes to 0 as i goes to infinity. And all in all, this implies the probability of Z_i minus 0 absolute value greater equal to epsilon in the limit of i going to infinity converges to 0 for all positive epsilon. And that completes our proof that indeed, Z_i converges to 0 in probability.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013

Transcript – Lecture 19

The following content is provided under a Creative Commons license. Your support will help MIT OpenCourseWare continue to offer high quality educational resources for free. To make a donation or view additional materials from hundreds of MIT courses, visit MIT OpenCourseWare at ocw.mit.edu.

JOHN TSITSIKLIS: We're going to start today a new unit. so we will be talking about limit theorems. So just to introduce the topic, let's think of the following situation. There's a population of penguins down at the South Pole. And if you were to pick a penguin at random and measure their height, the expected value of their height would be the average of the heights of the different penguins in the population. So suppose when you pick one, every penguin is equally likely. Then the expected value is just the average of all the penguins out there.

So your boss asks you to find out what that the expected value is. One way would be to go and measure each and every penguin. That might be a little time consuming. So alternatively, what you can do is to go and pick penguins at random, pick a few of them, let's say a number n of them. So you measure the height of each one. And then you calculate the average of the heights of those penguins that you have collected. So this is your estimate of the expected value.

Now, we called this the sample mean, which is the mean value, but within the sample that you have collected. This is something that's sort of feels the same as the expected value, which is again, the mean. But the expected value's a different kind of mean. The expected value is the mean over the entire population, whereas the sample mean is the average over the smaller sample that you have measured.

The expected value is a number. The sample mean is a random variable. It's a random variable because the sample you have collected is random.

Now, we think that this is a reasonable way of estimating the expectation. So in the limit as n goes to infinity, it's plausible that the sample mean, the estimate that we are constructing, should somehow get close to the expected value. What does this mean? What does it mean to get close? In what sense? And is this statement true?

This is the kind of statement that we deal with when dealing with limit theorems. That's the subject of limit theorems, when what happens if you're dealing with lots and lots of random variables, and perhaps take averages and so on.

So why do we bother about this? Well, if you're in the sampling business, it would be reassuring to know that this particular way of estimating the expected value actually gets you close to the true answer. There's also a higher level reason, which is a little more abstract and mathematical. So probability problems are easy to deal with if you're having in your hands one or two random

variables. You can write down their mass functions, joints density functions, and so on. You can calculate on paper or on a computer, you can get the answers.

Probability problems become computationally intractable if you're dealing, let's say, with 100 random variables and you're trying to get the exact answers for anything. So in principle, the same formulas that we have, they still apply. But they involve summations over large ranges of combinations of indices. And that makes life extremely difficult.

But when you push the envelope and you go to a situation where you're dealing with a very, very large number of variables, then you can start taking limits. And when you take limits, wonderful things happen. Many formulas start simplifying, and you can actually get useful answers by considering those limits. And that's sort of the big reason why looking at limit theorems is a useful thing to do.

So what we're going to do today, first we're going to start with a useful, simple tool that allows us to relates probabilities with expected values. The Markov inequality is the first inequality we're going to write down. And then using that, we're going to get the Chebyshev's inequality, a related inequality.

Then we need to define what do we mean by convergence when we talk about random variables. It's a notion that's a generalization of the notion of the usual convergence of limits of a sequence of numbers. And once we have our notion of convergence, we're going to see that, indeed, the sample mean converges to the true mean, converges to the expected value of the X's. And this statement is called the weak law of large numbers.

The reason it's called the weak law is because there's also a strong law, which is a statement with the same flavor, but with a somewhat different mathematical content. But it's a little more abstract, and we will not be getting into this. So the weak law is all that you're going to get.

All right. So now we start our digression. And our first tool will be the so-called Markov inequality. So let's take a random variable that's always non-negative. No matter what, it gets no negative values. To keep things simple, let's assume it's a discrete random variable. So the expected value is the sum over all possible values that a random variable can take. The values of the random variables that can take weighted according to their corresponding probabilities.

Now, this is a sum over all x's. But x takes non-negative values. And the PMF is also non-negative. So if I take a sum over fewer things, I'm going to get a smaller value. So the sum when I add over everything is less than or equal to the sum that I will get if I only add those terms that are bigger than a certain constant.

Now, if I'm adding over x's that are bigger than a, the x that shows up up there will always be larger than or equal to a. So we get this inequality. And now, a is a constant. I can pull it outside the summation. And then I'm left with the probabilities of all the x's that are bigger than a. And that's just the probability of being bigger than a.

OK, so that's the Markov inequality. Basically tells us that the expected value is larger than or equal to this number. It relates expected values to probabilities. It tells us that if the expected value is small, then the probability that x is big is also going to be small. So it's translates a statement about smallness of expected values to a statement about smallness of probabilities.

OK. What we actually need is a somewhat different version of this same statement. And what we're going to do is to apply this inequality to a non-negative random variable of a special type. And you can think of applying this same calculation to a random variable of this form, $(X - \mu)$ -squared, where μ is the expected value of X .

Now, this is a non-negative random variable. So, the expected value of this random variable, which is the variance, by following the same thinking as we had in that derivation up to there, is bigger than the probability that this random variable is bigger than some-- let me use a -squared instead of a times the value a -squared.

So now of course, this probability is the same as the probability that the absolute value of X minus μ is bigger than a -squared. And this side is equal to the variance of X . So this relates the variance of X to the probability that our random variable is far away from its mean. If the variance is small, then it means that the probability of being far away from the mean is also small.

So I derived this by applying the Markov inequality to this particular non-negative random variable. Or just to reinforce, perhaps, the message, and increase your confidence in this inequality, let's just look at the derivation once more, where I'm going, here, to start from first principles, but use the same idea as the one that was used in the proof out here.

Ok. So just for variety, now let's think of X as being a continuous random variable. The derivation is the same whether it's discrete or continuous. So by definition, the variance is the integral, is this particular integral. Now, the integral is going to become smaller if I integrate, instead of integrating over the full range, I only integrate over x 's that are far away from the mean. So μ is the mean. Think of c as some big number.

These are x 's that are far away from the mean to the left, from minus infinity to $\mu - c$. And these are the x 's that are far away from the mean on the positive side. So by integrating over fewer stuff, I'm getting a smaller integral.

Now, for any x in this range, this distance, $x - \mu$, is at least c . So that squared is at least c squared. So this term over this range of integration is at least c squared. So I can take it outside the integral. And I'm left just with the integral of the density. Same thing on the other side.

And so what factors out is this term c squared. And inside, we're left with the probability of being to the left of $\mu - c$, and then the probability of being to the right of $\mu + c$, which is the same as the probability that the absolute value of the distance from the mean is larger than or equal to c . So that's the same inequality that we proved there, except that here I'm using c . There I used a , but it's exactly the same one.

This inequality was maybe better to understand if you take that term and send it to the other side and write it this form. What does it tell us? It tells us that if c is a big number, it tells us that the probability of being more than c away from the mean is going to be a small number. When c is big, this is small.

Now, this is intuitive. The variance is a measure of the spread of the distribution, how wide it is. It tells us that if the variance is small, the distribution is not very wide. And mathematically, this translates to this statement that when the variance is small, the probability of being far away is going to be small. And the further away you're looking, that is, if c is a bigger number, that probability also becomes small.

Maybe an even more intuitive way to think about the content of this inequality is to, instead of c , use the number k , where k is positive and σ is the standard deviation. So let's just plug $k\sigma$ in the place of c . So this becomes $k\sigma^2$. These σ^2 's cancel. We're left with $1/k^2$.

Now, what is this? This is the event that you are k standard deviations away from the mean. So for example, this statement here tells you that if you look at the test scores from a quiz, what fraction of the class are 3 standard deviations away from the mean? It's possible, but it's not going to be a lot of people. It's going to be at most, $1/9$ of the class that can be 3 standard deviations or more away from the mean.

So the Chebyshev inequality is a really useful one. It comes in handy whenever you want to relate probabilities and expected values. So if you know that your expected values or, in particular, that your variance is small, this tells you something about tailed probabilities.

So this is the end of our first digression. We have this inequality in our hands. Our second digression is talk about limits. We want to eventually talk about limits of random variables, but as a warm up, we're going to start with limits of sequences.

So you're given a sequence of numbers, a_1, a_2, a_3 , and so on. And we want to define the notion that a sequence converges to a number. You sort of know what this means, but let's just go through it some more. So here's a . We have our sequence of values as n increases.

What do we mean by the sequence converging to a is that when you look at those values, they get closer and closer to a . So this value here is your typical a_{n_k} . They get closer and closer to a , and they stay closer. So let's try to make that more precise.

What it means is let's fix a sense of what it means to be close. Let me look at an interval that goes from $a - \epsilon$ to $a + \epsilon$. Then if my sequence converges to a , this means that as n increases, eventually the values of the sequence that I get stay inside this band. Since they converge to a , this means that eventually they will be smaller than $a + \epsilon$ and bigger than $a - \epsilon$.

So convergence means that given a band of positive length around the number a , the values of the sequence that you get eventually get inside and stay inside that band. So that's sort of the

picture definition of what convergence means. So now let's translate this into a mathematical statement.

Given a band of positive length, no matter how wide that band is or how narrow it is, so for every epsilon positive, eventually the sequence gets inside the band. What does eventually mean? There exists a time, so that after that time something happens. And the something that happens is that after that time, we are inside that band.

So this is a formal mathematical definition, which actually translates what I was telling in the wordy way before, and showing in terms of the picture. Given a certain band, even if it's narrow, eventually, after a certain time n_0 , the values of the sequence are going to stay inside this band.

Now, if I were to take epsilon to be very small, this thing would still be true that eventually I'm going to get inside of the band, except that I may have to wait longer for the values to get inside here. All right, that's what it means for a deterministic sequence to converge to something.

Now, how about random variables. What does it mean for a sequence of random variables to converge to a number? We're just going to twist a little bit of the word definition.

For numbers, we said that eventually the numbers get inside that band. But if instead of numbers we have random variables with a certain distribution, so here instead of a_n we're dealing with a random variable that has a distribution, let's say, of this kind, what we want is that this distribution gets inside this band, so it gets concentrated inside here. What does it mean that the distribution gets inside this band?

I mean a random variable has a distribution. It may have some tails, so maybe not the entire distribution gets concentrated inside of the band. But we want that more and more of this distribution is concentrated in this band. So that -- in a sense that -- the probability of falling outside the band converges to 0 -- becomes smaller and smaller.

So in words, we're going to say that the sequence random variables or a sequence of probability distributions, that would be the same, converges to a particular number a if the following is true. If I consider a small band around a , then the probability that my random variable falls outside this band, which is the area under this curve, this probability becomes smaller and smaller as n goes to infinity. The probability of being outside this band converges to 0. So that's the intuitive idea.

So in the beginning, maybe our distribution is sitting everywhere. As n increases, the distribution starts to get concentrating inside the band. When a is even bigger, our distribution is even more inside that band, so that these outside probabilities become smaller and smaller.

So the corresponding mathematical statement is the following. I fix a band around a , a +/- epsilon. Given that band, the probability of falling outside this band, this probability converges to 0. Or another way to say it is that the limit of this probability is equal to 0.

If you were to translate this into a complete mathematical statement, you would have to write down the following messy thing. For every epsilon positive -- that's this statement -- the limit is 0.

What does it mean that the limit of something is 0? We flip back to the previous slide. Why? Because a probability is a number. So here we're talking about a sequence of numbers convergent to 0.

What does it mean for a sequence of numbers to converge to 0? It means that for any epsilon prime positive, there exists some n_0 such that for every n bigger than n_0 the following is true -- that this probability is less than or equal to epsilon prime.

So the mathematical statement is a little hard to parse. For every size of that band, and then you take the definition of what it means for the limit of a sequence of numbers to converge to 0. But it's a lot easier to describe this in words and, basically, think in terms of this picture. That as n increases, the probability of falling outside those bands just become smaller and smaller. So the statement is that our distribution gets concentrated in arbitrarily narrow little bands around that particular number a .

OK. So let's look at an example. Suppose a random variable Y_n has a discrete distribution of this particular type. Does it converge to something? Well, the probability distribution of this random variable gets concentrated at 0 -- there's more and more probability of being at 0.

If I fix a band around 0 -- so if I take the band from minus epsilon to epsilon and look at that band-- the probability of falling outside this band is $1/n$. As n goes to infinity, that probability goes to 0. So in this case, we do have convergence. And Y_n converges in probability to the number 0. So this just captures the facts obvious from this picture, that more and more of our probability distribution gets concentrated around 0, as n goes to infinity.

Now, an interesting thing to notice is the following, that even though Y_n converges to 0, if you were to write down the expected value for Y_n , what would it be? It's going to be n times the probability of this value, which is $1/n$. So the expected value turns out to be 1. And if you were to look at the expected value of Y_n -squared, this would be 0. times this probability, and then n -squared times this probability, which is equal to n . And this actually goes to infinity.

So we have this, perhaps, strange situation where a random variable goes to 0, but the expected value of this random variable does not go to 0. And the second moment of that random variable actually goes to infinity. So this tells us that convergence in probability tells you something, but it doesn't tell you the whole story. Convergence to 0 of a random variable doesn't imply anything about convergence of expected values or of variances and so on.

So the reason is that convergence in probability tells you that this tail probability here is very small. But it doesn't tell you how far does this tail go. As in this example, the tail probability is small, but that tail acts far away, so it gives a disproportionate contribution to the expected value or the expected value squared.

OK. So now we've got everything that we need to go back to the sample mean and study its properties. So the sad thing is that we have a sequence of random variables. They're independent. They have the same distribution. And we assume that they have a finite mean and a finite variance. We're looking at the sample mean.

Now in principle, you can calculate the probability distribution of the sample mean, because we know how to find the distributions of sums of independent random variables. You use the convolution formula over and over. But this is pretty complicated, so let's not look at that. Let's just look at expected values, variances, and the probabilities that the sample mean is far away from the true mean.

So what is the expected value of this random variable? The expected value of a sum of random variables is the sum of the expected values. And then we have this factor of n in the denominator. Each one of these expected values is μ , so we get μ . So the sample mean, the average value of this M_n in expectation is the same as the true mean inside our population.

Now here, this is a fine conceptual point, there's two kinds of averages involved when you write down this expression. We understand that expectations are some kind of average. The sample mean is also an average over the values that we have observed.

But it's two different kinds of averages. The sample mean is the average of the heights of the penguins that we collected over a single expedition. The expected value is to be thought of as follows, my probabilistic experiment is one expedition to the South Pole. Expected value here means thinking on the average over a huge number of expeditions.

So my expedition is a random experiment, I collect random samples, and they record M_n . The average result of an expedition is what we would get if we were to carry out a zillion expeditions and average the averages that we get at each particular expedition. So this M_n is the average during a single expedition. This expectation is the average over an imagined infinite sequence of expeditions. And of course, the other thing to always keep in mind is that expectations give you numbers, whereas the sample mean is actually a random variable.

All right. So this random variable, how random is it? How big is its variance? So the variance of a sum of random variables is the sum of the variances. But since we're dividing by n , when you calculate variances this brings in a factor of n -squared. So the variance is σ^2/n .

And in particular, the variance of the sample mean becomes smaller and smaller. It means that when you estimate that average height of penguins, if you take a large sample, then your estimate is not going to be too random. The randomness in your estimates become small if you have a large sample size. Having a large sample size kind of removes the randomness from your experiment.

Now let's apply the Chebyshev inequality to say something about tail probabilities for the sample mean. The probability that you are more than ϵ away from the true mean is less than or equal to the variance of this quantity divided by this number squared. So that's just the translation

of the Chebyshev inequality to the particular context we've got here. We found the variance. It's sigma-squared over n. So we end up with this expression.

So what does this expression do? For any given epsilon, if I fix epsilon, then this probability, which is less than sigma-squared over n epsilon-squared, converges to 0 as n goes to infinity. And this is just the definition of convergence in probability. If this happens, that the probability of being more than epsilon away from the mean, that probability goes to 0, and this is true no matter how I choose my epsilon, then by definition we have convergence in probability.

So we have proved that the sample mean converges in probability to the true mean. And this is what the weak law of large numbers tells us. So in some vague sense, it tells us that the sample means, when you take the average of many, many measurements in your sample, then the sample mean is a good estimate of the true mean in the sense that it approaches the true mean as your sample size increases. It approaches the true mean, but of course in a very specific sense, in probability, according to this notion of convergence that we have used.

So since we're talking about sampling, let's go over an example, which is the typical situation faced by someone who's constructing a poll. So you're interested in some property of the population. So what fraction of the population prefers Coke to Pepsi? So there's a number f, which is that fraction of the population. And so this is an exact number. So out of a population of 100 million, 20 million prefer Coke, then f would be 0.2.

We want to find out what that fraction is. We cannot ask everyone. What we're going to do is to take a random sample of people and ask them for their preferences. So the ith person either says yes for Coke or no. And we record that by putting a 1 each time that we get a yes answer.

And then we form the average of these x's. What is this average? It's the number of 1's that we got divided by n. So this is a fraction, but calculated only on the basis of the sample that we have. So you can think of this as being an estimate, f_{hat} , based on the sample that we have.

Now, even though we used the lower case letter here, this f_{hat} is, of course, a random variable. f is a number. This is the true fraction in the overall population. f_{hat} is the estimate that we get by using our particular sample.

Ok. So your boss told you, I need to know what f is, but go and do some sampling. What are you going to respond? Unless I ask everyone in the whole population, there's no way for me to know f exactly. Right? There's no way.

OK, so the boss tells you, well OK, then that'll me f within an accuracy. I want an answer from you, that's your answer, which is close to the correct answer within 1 % point. So if the true f is 0.4, your answer should be somewhere between 0.39 and 0.41. I want a really accurate answer.

What are you going to say? Well, there's no guarantee that my answer will be within 1 %. Maybe I'm unlucky and I just happen to sample the wrong set of people and my answer comes out to be wrong. So I cannot give you a hard guarantee that this inequality will be satisfied.

But perhaps, I can give you a guarantee that this inequality will be satisfied, this accuracy requirement will be satisfied, with high confidence. That is, there's going to be a smaller probability that things go wrong, that I'm unlikely and I use a bad sample. But leaving aside that smaller probability of being unlucky, my answer will be accurate within the accuracy requirement that you have.

So these two numbers are the usual specs that one has when designing polls. So this number is the accuracy that we want. It's the desired accuracy. And this number has to do with the confidence that we want. So 1 minus that number, we could call it the confidence that we want out of our sample. So this is really 1 minus confidence.

So now your job is to figure out how large an n , how large a sample should you be using, in order to satisfy the specs that your boss gave you. All you know at this stage is the Chebyshev inequality. So you just try to use it. The probability of getting an answer that's more than 0.01 away from the true answer is, by Chebyshev's inequality, the variance of this random variable divided by this number squared. The variance, as we argued a little earlier, is the variance of the x 's divided by n . So we get this expression. So we would like this number to be less than or equal to 0.05.

OK, here we hit a little bit off a difficulty. The variance, $(\sigma_x)^2$, what is it? $(\sigma_x)^2$ is, if you remember the variance of a Bernoulli random variable, is this quantity. But we don't know it. f is what we're trying to estimate in the first place. So the variance is not known, so I cannot plug in a number inside here.

What I can do is to be conservative and use an upper bound of the variance. How large can this number get? Well, you can plot f times $(1-f)$. It's a parabola. It has a root at 0 and at 1. So the maximum value is going to be, by symmetry, at $1/2$ and when f is $1/2$, then this variance becomes $1/4$.

So I don't know $(\sigma_x)^2$, but I'm going to use the worst case value for $(\sigma_x)^2$, which is 4. And this is now an inequality that I know to be always true. I've got my specs, and my specs tell me that I want this number to be less than 0.05.

And given what I know, the best thing I can do is to say, OK, I'm going to take this number and make it less than 0.05. If I choose my n so that this is less than 0.05, then I'm certain that this probability is also less than 0.05.

What does it take for this inequality to be true? You can solve for n here, and you find that to satisfy this inequality, n should be larger than or equal to 50,000. So you can just let n be equal to 50,000. So the Chebyshev inequality tells us that if you take n equal to 50,000, then by the Chebyshev inequality, we're guaranteed to satisfy the specs that we were given.

Ok. Now, 50,000 is a bit of a large sample size. Right? If you read anything in the newspapers where they say so much of the voters think this and that, this was determined on the basis of a sample of 1,200 likely voters or so. So the numbers that you will typically see in these news

items about polling, they usually involve sample sizes about the 1,000 or so. You will never see a sample size of 50,000. That's too much.

So where can we cut some corners? Well, we can cut corners basically in three places. This requirement is a little too tight. Newspaper stories will usually tell you, we have an accuracy of $\pm 3\%$ points, instead of 1% point. And because this number comes up as a square, by making it 3% points instead of 1, saves you a factor of 10.

Then, the five percent confidence, I guess that's usually OK. If we use that factor of 10, then we make our sample that we gain from here, then we get a sample size of 10,000. And that's, again, a little too big. So where can we fix things?

Well, it turns out that this inequality that we're using here, Chebyshev's inequality, is just an inequality. It's not that tight. It's not very accurate. Maybe there's a better way of calculating or estimating this quantity, which is smaller than this. And using a more accurate inequality or a more accurate bound, then we can convince ourselves that we can settle with a smaller sample size.

This more accurate kind of inequality comes out of a difference limit theorem, which is the next limit theorem we're going to consider. We're going to start the discussion today, but we're going to continue with it next week.

Before I tell you exactly what that other limit theorem says, let me give you the big picture of what's involved here. We're dealing with sums of i.i.d random variables. Each X has a distribution of its own.

So suppose that X has a distribution which is something like this. This is the density of X . If I add lots of X 's together, what kind of distribution do I expect? The mean is going to be n times the mean of an individual X . So if this is μ , I'm going to get a mean of n times μ .

But my variance will also increase. When I add the random variables, I'm adding the variances. So since the variance increases, we're going to get a distribution that's pretty wide. So this is the density of X_1 plus all the way to X_n . So as n increases, my distribution shifts, because the mean is positive. So I keep adding things. And also, my distribution becomes wider and wider. The variance increases.

Well, we started a different scaling. We started a scaled version of this quantity when we looked at the weak law of large numbers. In the weak law of large numbers, we take this random variable and divide it by n . And what the weak law tells us is that we're going to get a distribution that's very highly concentrated around the true mean, which is μ .

So this here would be the density of X_1 plus X_n divided by n . Because I've divided by n , the mean has become the original mean, which is μ . But the weak law of large numbers tells us that the distribution of this random variable is very concentrated around the mean. So we get a distribution that's very narrow in this kind. In the limit, this distribution becomes one that's just concentrated on top of μ . So it's sort of a degenerate distribution.

So these are two extremes, no scaling for the sum, a scaling where we divide by n . In this extreme, we get the trivial case of a distribution that flattens out completely. In this scaling, we get a distribution that gets concentrated around a single point.

Again, we look at some intermediate scaling that makes things more interesting. Things do become interesting if we scale by dividing the sum by square root of n instead of dividing by n . What effect does this have?

When we scale by dividing by square root of n , the variance of S_n over square root of n is going to be the variance of S_n over sum divided by n . That's how variances behave. The variance of S_n is $n \sigma^2$, divide by n , which is σ^2 , which means that when we scale in this particular way, as n changes, the variance doesn't change.

So the width of our distribution will be sort of constant. The distribution changes shape, but it doesn't become narrower as was the case here. It doesn't become wider, kind of keeps the same width. So perhaps in the limit, this distribution is going to take an interesting shape. And that's indeed the case.

So let's do what we did before. So we're looking at the sum, and we want to divide the sum by something that goes like square root of n . So the variance of S_n is $n \sigma^2$. The variance of the standard deviation of S_n is the square root of that. It's this number. So effectively, we're scaling by order of square root n .

Now, I'm doing another thing here. If my random variable has a positive mean, then this quantity is going to have a mean that's positive and growing. It's going to be shifting to the right.

Why is that? S_n has a mean that's proportional to n . When I divide by square root n , then it means that the mean scales like square root of n . So my distribution would still keep shifting after I do this division.

I want to keep my distribution in place, so I subtract out the mean of S_n . So what we're doing here is a standard technique or transformation where you take a random variable and you so-called standardize it. I remove the mean of that random variable and I divide by the standard deviation. This results in a random variable that has 0 mean and unit variance.

What Z_n measures is the following, Z_n tells me how many standard deviations am I away from the mean. S_n minus (n times expected value of X) tells me how much is S_n away from the mean value of S_n . And by dividing by the standard deviation of S_n -- this tells me how many standard deviations away from the mean am I.

So we're going to look at this random variable, which is just a transformation Z_n . It's a linear transformation of S_n . And we're going to compare this random variable to a standard normal random variable.

So a standard normal is the random variable that you are familiar with, given by the usual formula, and for which we have tables for it. This Z_n has 0 mean and unit variance. So in that

respect, it has the same statistics as the standard normal. The distribution of Z_n could be anything -- can be pretty messy.

But there is this amazing theorem called the central limit theorem that tells us that the distribution of Z_n approaches the distribution of the standard normal in the following sense, that probability is that you can calculate -- of this type -- that you can calculate for Z_n -- is the limit becomes the same as the probabilities that you would get from the standard normal tables for Z .

It's a statement about the cumulative distribution functions. This quantity, as a function of c , is the cumulative distribution function of the random variable Z_n . This is the cumulative distribution function of the standard normal. The central limit theorem tells us that the cumulative distribution function of the sum of a number of random variables, after they're appropriately standardized, approaches the cumulative distribution function over the standard normal distribution.

In particular, this tells us that we can calculate probabilities for Z_n when n is large by calculating instead probabilities for Z . And that's going to be a good approximation. Probabilities for Z are easy to calculate because they're well tabulated. So we get a very nice shortcut for calculating probabilities for Z_n .

Now, it's not Z_n that you're interested in. What you're interested in is S_n . And S_n -- inverting this relation here -- S_n is square root n sigma Z_n plus n expected value of X .

All right. Now, if you can calculate probabilities for Z_n , even approximately, then you can certainly calculate probabilities for S_n , because one is a linear function of the other. And we're going to do a little bit of that next time. You're going to get, also, some practice in recitation. At a more vague level, you could describe the central limit theorem as saying the following, when n is large, you can pretend that Z_n is a standard normal random variable and do the calculations as if Z_n was standard normal.

Now, pretending that Z_n is normal is the same as pretending that S_n is normal, because S_n is a linear function of Z_n . And we know that linear functions of normal random variables are normal. So the central limit theorem essentially tells us that we can pretend that S_n is a normal random variable and do the calculations just as if it were a normal random variable.

Mathematically speaking though, the central limit theorem does not talk about the distribution of S_n , because the distribution of S_n becomes degenerate in the limit, just a very flat and long thing. So strictly speaking mathematically, it's a statement about cumulative distributions of Z_n 's. Practically, the way you use it is by just pretending that S_n is normal.

Very good. Enjoy the Thanksgiving Holiday.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Problem Set 9
Due November 22, 2010

1. Random variable X is uniformly distributed between -1.0 and 1.0 . Let X_1, X_2, \dots be independent identically distributed random variables with the same distribution as X . Determine which, if any, of the following sequences (all with $i = 1, 2, \dots$) are convergent in probability. Fully justify your answers. Include the limits if they exist.

$$(a) U_i = \frac{X_1 + X_2 + \dots + X_i}{i}$$

$$(b) W_i = \max(X_1, \dots, X_i)$$

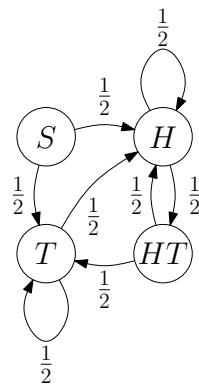
$$(c) V_i = X_1 \cdot X_2 \cdot \dots \cdot X_i$$

2. Demonstrate that the Chebyshev inequality is tight, that is, for every $\mu, \sigma > 0$, and $c \geq \sigma$, construct a random variable X with mean μ and standard deviation σ such that

$$\mathbf{P}(|X - \mu| \geq c) = \frac{\sigma^2}{c^2}$$

Hint: You should be able to do this with a discrete random variable that takes on only 3 distinct values with nonzero probability.

3. Assume that a fair coin is tossed repeatedly, with the tosses being independent. We want to determine the expected number of tosses necessary to first observe a head directly followed by a tail. To do so, we define a Markov chain with states S, H, T, HT , where S is a starting state, H indicates a head on the current toss, T indicates a tail on the current toss (without heads on the previous toss), and HT indicates heads followed by tails over the last two tosses. This Markov chain is illustrated below:



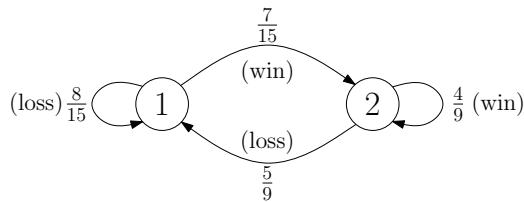
We can find the expected number of tosses necessary to first observe a heads directly followed by tails by solving a mean first passage time problem for this Markov chain.

- (a) What is the expected number of tosses necessary to first observe a head directly followed by tails?
- (b) Assuming we have just observed a head followed by a tail, what is the expected number of additional tosses until we again observe a head followed directly by a tail?

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
 (Fall 2010)

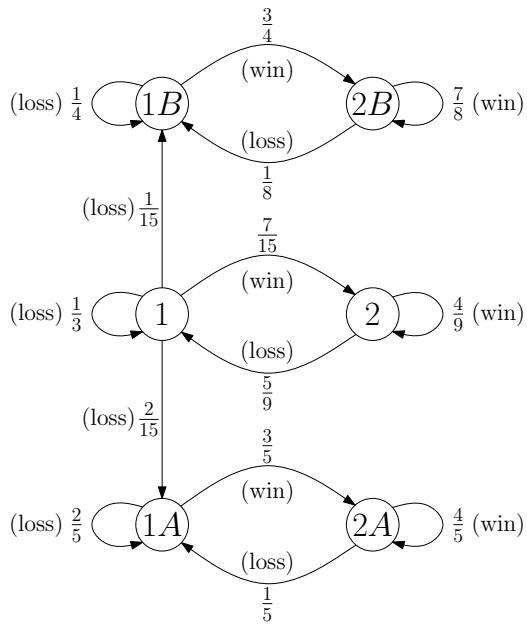
Next, we want to answer the same questions for the event tails directly followed by tails. Set up a different Markov chain from which we could calculate the expected number of tosses necessary to first observe tails directly followed by tails.

- (c) What is the expected number of tosses necessary to first observe a tail directly followed by a tail?
 - (d) Assuming we have just observed a tail followed by a tail, what is the expected number of additional tosses until we again observe a tail followed directly by a tail? Note that the number of additional tosses could be as little as one, if tails were to come up again.
4. Jack is a gambler who pays for his MIT tuition by spending weekends in Las Vegas. Lately he's been playing 21 at a table that returns cards to the deck and reshuffles them all before each hand. As he has a fixed policy in how he plays, his probability of winning a particular hand remains constant, and is independent of all other hands. There is a wrinkle, however; the dealer switches between two decks (deck #2 is more unfair to Jack than deck #1), depending on whether or not Jack wins. Jack's wins and losses can be modeled via the transitions of the following Markov chain, whose states correspond to the particular deck being used.



- (a) What is Jack's long term probability of winning?

Given that Jack loses and the dealer is not occupied with switching decks, with probability $\frac{2}{8}$ the dealer looks away for one second and with probability $\frac{1}{8}$ the dealer looks away for two seconds, independently of everything else. When this happens, Jack secretly inserts additional cards into both of the dealer's decks, transforming the decks into types 1A & 2A (when he has 1 second) or 1B & 2B (when he has 2 seconds). Jack slips cards into the decks at most once. The process can be described by the modified Markov chain in the picture. Assume in all future problems that play begins with the dealer using deck #1.



- (b) What is the probability of Jack eventually playing with decks 1A and 2A?
- (c) What is Jack's long-term probability of winning?
- (d) What is the expected time (as in number of hands) until Jack slips additional cards into the deck?
- (e) What is the distribution of the number of times that the dealer switches from deck 2 to deck 1?
- (f) What is the distribution of the number of wins that Jack has before he slips extra cards into the deck? *Hint:* Note that after some conditioning, we have a geometric number of geometric random variables, all of which are independent.
- (g) What is the average net losses (number of losses minus the number of wins, sometimes negative) prior to Jack slipping additional cards into the deck?
- (h) Given that after a very long period of time Jack is playing a hand with deck 1A, what is the approximate probability that his previous hand was played with deck 2A?

G1[†]. Show the following one-sided version of Chebyshev's inequality:

$$\mathbf{P}(X - \mu \geq a) \leq \frac{\sigma^2}{(\sigma^2 + a^2)}$$

where μ and σ^2 are the mean and variance of X respectively, and $a > 0$. Hint: Start by finding a bound on $\mathbf{P}(X - \mu + c \geq a + c)$ with $c \geq 0$. Then find the c that 'tightens' your bound.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Problem Set 9 Solutions

1. (a) Yes, to 0. Applying the weak law of large numbers, we have

$$\mathbf{P}(|U_i - \mu| > \epsilon) \rightarrow 0 \text{ as } i \rightarrow \infty, \text{ for all } \epsilon > 0$$

Here $\mu = 0$ since $X_i \sim U(-1.0, 1.0)$.

- (b) Yes, to 1. Since $W_i \leq 1$, we have for $\epsilon > 0$,

$$\begin{aligned} \lim_{i \rightarrow \infty} \mathbf{P}(|W_i - 1| > \epsilon) &= \lim_{i \rightarrow \infty} \mathbf{P}(\max\{X_1, \dots, X_i\} < 1 - \epsilon) \\ &= \lim_{i \rightarrow \infty} \mathbf{P}(X_1 < 1 - \epsilon) \cdots \mathbf{P}(X_i < 1 - \epsilon) \\ &= \lim_{i \rightarrow \infty} (1 - \frac{\epsilon}{2})^i \\ &= 0. \end{aligned}$$

- (c) Yes, to 0.

$$|V_n| \leq \min\{|X_1|, |X_2|, \dots, |X_n|\}$$

but $\min\{|X_1|, |X_2|, \dots, |X_n|\}$ converges to 0 in probability. So, since $|V_n| \geq 0$, $|V_n|$ converges to 0 in probability. To see why $\min\{|X_1|, |X_2|, \dots, |X_n|\}$ converges to 0 in probability, note that:

$$\begin{aligned} \lim_{i \rightarrow \infty} \mathbf{P}(|\min\{|X_1|, \dots, |X_i|\} - 0| > \epsilon) &= \lim_{i \rightarrow \infty} \mathbf{P}(\min\{|X_1|, \dots, |X_i|\} > \epsilon) \\ &= \lim_{i \rightarrow \infty} \mathbf{P}(|X_1| > \epsilon) \cdot \mathbf{P}(|X_2| > \epsilon) \cdots \mathbf{P}(|X_i| > \epsilon) \\ &= \lim_{i \rightarrow \infty} (1 + \epsilon)^i \text{ since } |X_i| \text{ is uniform between 0 and 1} \\ &= 0. \end{aligned}$$

2. Consider a random variable X with PMF

$$p_X(x) = \begin{cases} p, & \text{if } x = \mu - c; \\ p, & \text{if } x = \mu + c; \\ 1 - 2p, & \text{if } x = \mu. \end{cases}$$

The mean of X is μ , and the variance of X is $2pc^2$. To make the variance equal σ^2 , set $p = \frac{\sigma^2}{2c^2}$. For this random variable, we have

$$\mathbf{P}(|X - \mu| \geq c) = 2p = \frac{\sigma^2}{c^2},$$

and therefore the Chebyshev inequality is tight.

3. (a) Let t_i be the expected time until the state HT is reached, starting in state i , i.e., the mean first passage time to reach state HT starting in state i . Note that t_S is the expected number of tosses until first observing heads directly followed by tails. We have,

$$\begin{aligned} t_S &= 1 + \frac{1}{2}t_H + \frac{1}{2}t_T \\ t_T &= 1 + \frac{1}{2}t_H + \frac{1}{2}t_T \\ t_H &= 1 + \frac{1}{2}t_H \end{aligned}$$

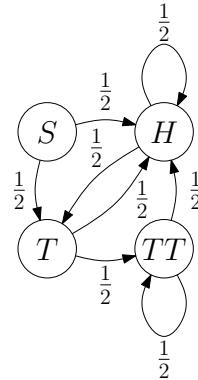
and by solving these equations, we find that the expected number of tosses until first observing heads directly followed by tails is

$$t_S = 4.$$

- (b) To find the expected number of additional tosses necessary to again observe heads followed by tails, we recognize that this is the mean recurrence time t_{HT}^* of state HT . This can be determined as

$$\begin{aligned} t_{HT}^* &= 1 + p_{HT,H}t_H + p_{HT,T}t_T \\ &= 1 + \frac{1}{2} \cdot 2 + \frac{1}{2} \cdot 4 \\ &= 4. \end{aligned}$$

- (c) Let's consider a Markov chain with states S, H, T, TT , where S is a starting state, H indicates heads on the current toss, T indicates tails on the current toss (without tails on the previous toss), and TT indicates tails over the last two tosses. The transition probabilities for this Markov chain are illustrated below in the state transition diagram:



Let t_i be the expected time until the state TT is reached, starting in state i , i.e., the mean first passage time to reach state TT starting in state i . Note that t_S is the expected number of tosses until first observing tails directly followed by tails. We have,

$$\begin{aligned} t_S &= 1 + \frac{1}{2}t_H + \frac{1}{2}t_T \\ t_T &= 1 + \frac{1}{2}t_H \\ t_H &= 1 + \frac{1}{2}t_H + \frac{1}{2}t_T \end{aligned}$$

and by solving these equations, we find that the expected number of tosses until first observing two consecutive tails is

$$t_S = 6.$$

- (d) To find the expected number of additional tosses necessary to again observe heads followed by tails, we recognize that this is the mean recurrence time t_{TT}^* of state TT . This can be

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
 (Fall 2010)

determined as

$$\begin{aligned} t_{TT}^* &= 1 + p_{TT,HT}t_H + p_{TT,TT}t_{TT} \\ &= 1 + \frac{1}{2} \cdot 6 + \frac{1}{2} \cdot 0 \\ &= 4. \end{aligned}$$

It may be surprising that the average number of tosses until the first two consecutive tails is greater than the average number of tosses until heads is directly followed by tails, considering that the mean recurrence time between pairs of tosses with heads directly followed by tails equals the mean recurrence time between pairs of tosses that are both tails (or equivalently, the long-term frequency of pairs of tosses with heads followed by tails equals the long-term frequency of pairs of tosses with two consecutive tails¹). This is a start-up artifact. Note that the distribution of the first passage time to reach state HT (or TT) starting in state S is the same as the conditional distribution of the recurrence time of state HT (or TT), given that it is greater than 1. Although in both cases the *expected values* of the recurrence times are equal (this is what parts (b) and (d) tell us), the conditional expected values of the recurrence time given that it is greater than 1 is not the same in both cases (possibly, because the unconditional distributions are not equal).

4. (a) The long-term frequency of winning can be found as sum of the long-term frequency of transitions from 1 to 2 and 2 to 2. These can be found from the steady-state probabilities π_1 and π_2 , which are known to exist as the chain is aperiodic and recurrent. The local balance and normalization equations are as follows:

$$\begin{aligned} \frac{7}{15}\pi_1 &= \frac{5}{9}\pi_2, \\ \pi_1 + \pi_2 &= 1. \end{aligned}$$

Solving these we obtain,

$$\pi_1 = \frac{25}{46} \approx 0.54, \quad \pi_2 = \frac{21}{46} \approx 0.46.$$

The probability of winning, which is the long-term frequency of the transitions from 1 to 2 and 2 to 2, can now be found as

$$P(\text{winning}) = \pi_1 p_{12} + \pi_2 p_{22} = \frac{25}{46} \frac{7}{15} + \frac{21}{46} \frac{4}{9} = \frac{21}{46} \approx 0.46.$$

Note that from the balance equation for state 2,

$$\pi_2 = \pi_1 p_{12} + \pi_2 p_{22},$$

the long-term probability of winning always equals π_2 .

- (b) This question is one of determining the probability of absorption into the recurrent class $\{1A, 2A\}$. This probability of absorption can be found by recognizing that it will be the ratio of probabilities

$$\frac{p_{1,1A}}{p_{1,1A} + p_{1,1B}} = \frac{\frac{2}{15}}{\frac{2}{15} + \frac{1}{15}} = \frac{2}{3}.$$

¹See problem 7.34 on page 399 of the text for a detailed explanation of this correspondence between mean recurrence times and steady-state probabilities.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
 (Fall 2010)

More methodically, if we define a_i as the probability of being absorbed into the class $\{1A, 2A\}$, starting in state i , we can solve for the a_i by solving the system of equations

$$\begin{aligned} a_1 &= p_{1,1A} + p_{11}a_1 + p_{12}a_2 \\ &= \frac{2}{15} + \frac{1}{3}a_1 + \frac{7}{15}a_2 \\ a_2 &= p_{21}a_1 + p_{22}a_2 \\ &= \frac{5}{9}a_1 + \frac{4}{9}a_2, \end{aligned}$$

from which we determine that $a_1 = \frac{p_{1,1A}}{p_{1,1A}+p_{1,1B}} = \frac{2}{3}$.

- (c) Let A, B be the events that Jack eventually plays with decks $1A$ & $2A$, $1B$ & $2B$, respectively, when starting in state 1. From part (b), we know that $\mathbf{P}(A) = a_1 = \frac{2}{3}$ and $\mathbf{P}(B) = 1 - a_1 = \frac{1}{3}$. The probability of winning can be determined as

$$\mathbf{P}(\text{winning}) = \mathbf{P}(\text{winning}|A)\mathbf{P}(A) + \mathbf{P}(\text{winning}|B)\mathbf{P}(B).$$

By considering the corresponding the appropriate recurrent class and solving a problem similar to part (a), $\mathbf{P}(\text{winning}|A)$ and $\mathbf{P}(\text{winning}|B)$ can be determined; in these cases, the steady-state probabilities of each recurrent class are defined under the assumption of being absorbed into that particular recurrent class. Let's begin with $\mathbf{P}(\text{winning}|A)$. The local balance and normalization equations for the recurrent class $\{1A, 2A\}$ are

$$\begin{aligned} \frac{3}{5}\pi_{1A} &= \frac{1}{5}\pi_{2A}, \\ \pi_{1A} + \pi_{2A} &= 1. \end{aligned}$$

Solving these we obtain,

$$\pi_{1A} = \frac{1}{4}, \quad \pi_{2A} = \frac{3}{4},$$

and hence conclude that

$$\mathbf{P}(\text{winning}|A) = p_{1A,2A}\pi_{1A} + p_{2A,2A}\pi_{2A} = \pi_{2A} = \frac{3}{4}.$$

Similarly, the local balance and normalization equations for the recurrent class $\{1B, 2B\}$ are

$$\begin{aligned} \frac{3}{4}\pi_{1B} &= \frac{1}{8}\pi_{2B}, \\ \pi_{1B} + \pi_{2B} &= 1. \end{aligned}$$

Solving these we obtain,

$$\pi_{1B} = \frac{1}{7}, \quad \pi_{2B} = \frac{6}{7},$$

and hence conclude that

$$\mathbf{P}(\text{winning}|B) = p_{1B,2B}\pi_{1B} + p_{2B,2B}\pi_{2B} = \pi_{2B} = \frac{6}{7}.$$

Putting these pieces together, we have that

$$\begin{aligned}\mathbf{P}(\text{winning}) &= \mathbf{P}(\text{winning}|A)\mathbf{P}(A) + \mathbf{P}(\text{winning}|B)\mathbf{P}(B) \\ &= \frac{3}{4} \cdot \frac{2}{3} + \frac{6}{7} \cdot \frac{1}{3} \\ &= \frac{11}{14} \approx 0.79 ,\end{aligned}$$

meaning that Jack substantially increases the odds to his favor by slipping additional cards into the decks.

- (d) The expected time until Jack slips cards into the deck is the same as the expected time until the Markov chain enters a recurrent state. Let μ_i be the expected amount of time until a recurrent state is reached from state i . We have the equations

$$\begin{aligned}\mu_1 &= 1 + p_{11}\mu_1 + p_{12}\mu_2 = 1 + \frac{1}{3}\mu_1 + \frac{7}{15}\mu_2 \\ \mu_2 &= 1 + p_{21}\mu_1 + p_{22}\mu_2 = 1 + \frac{5}{9}\mu_1 + \frac{4}{9}\mu_2 ,\end{aligned}$$

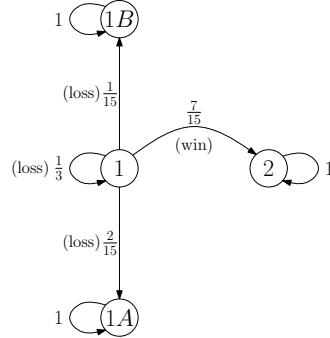
which when solved, yields the expected time until Jack slips cards into the deck,

$$\mu_1 = 9.2 .$$

- (e) Let S be the number of times that the dealer switches from deck #2 to deck #1, which equals the number of times that he/she switches from deck #1 to deck #2. Let p be the probability that $S = 0$, which is the sum of the probability of all ways for the first change of state to be from state 1 to state 1A or state 1B,

$$p = \left(\frac{2}{15} + \frac{1}{15} \right) + \left(\frac{1}{3} \right) \left(\frac{2}{15} + \frac{1}{15} \right) + \left(\frac{1}{3} \right)^2 \left(\frac{2}{15} + \frac{1}{15} \right) + \dots = \frac{1}{1 - 1/3} \cdot \frac{3}{15} = \frac{3}{10} .$$

Alternatively, p is the probability of absorption of the following modified chain into an absorbing state (1A or 1B), when started in state 1:



As $\mathbf{P}(S > 0) = 1 - p$, and similarly, $\mathbf{P}(S > k + 1|S > k) = 1 - p$, it should be clear that S will be a shifted geometric, and thus

$$p_S(k) = \left(\frac{7}{10} \right)^k \frac{3}{10} \quad k = 0, 1, 2, \dots .$$

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
 (Fall 2010)

- (f) Note that S from part (e) is the total number of cycles from 1 to 2 and back to 1. During the i th cycle, the number of wins, W_i , is a geometric random variable with parameter $q = \frac{5}{9}$. Thus the total number of wins by Jack before he slips extra cards into the deck is

$$W = W_1 + W_2 + \dots + W_S ,$$

which is a random number of random variables, all of which are independent. Conditioned on $S > 0$, W is a geometric (with parameter p) number of geometric (with parameter q) random variables, all conditionally independent, and thus from the theory of splitting Bernoulli processes,

$$p_{W|S>0}(k) = (1 - pq)^{k-1} pq \quad k = 1, 2, \dots ,$$

where $pq = \frac{3}{10} \cdot \frac{5}{9} = \frac{1}{6}$. When $S = 0$, it follows that $W = 0$, and thus by total probability,

$$p_W(k) = \begin{cases} \frac{3}{10} & k = 0 \\ (\frac{7}{10})(\frac{5}{6})^{k-1} \frac{1}{6} & k = 1, 2, \dots . \end{cases}$$

- (g) Let W be the total number of wins before slipping cards into the deck (as in part (f)), and similarly let L be the total number of losses before absorption. We know from part (d) that $\mathbf{E}[W + L] = \mu_1 = 9.2$. From part (f) we can find $\mathbf{E}[W]$ by total expectation,

$$\mathbf{E}[W] = E[W|S=0]\mathbf{P}(S=0) + E[W|S>0]\mathbf{P}(S>0) = \frac{7/10}{1/6} = \frac{42}{10} = 4.2 ,$$

because when conditioned on $S > 0$, the number of wins, W , is a geometric random variable with parameter $pq = \frac{1}{6}$. From linearity of expectation, we find

$$\mathbf{E}[L - W] = \mathbf{E}[W + L] - 2\mathbf{E}[W] = 9.2 - 2 \cdot 4.2 = 0.8 .$$

- (h) Using A to again denote the probability of being absorbed into the recurrent class $\{1A, 2A\}$, starting in state 1,

$$\begin{aligned} \mathbf{P}(X_n = 2A | X_{n+1} = 1A) &= \frac{\mathbf{P}(X_{n+1} = 1A | X_n = 2A)\mathbf{P}(X_n = 2A)}{\mathbf{P}(X_{n+1} = 1A)} \\ &= \frac{\mathbf{P}(X_{n+1} = 1A | X_n = 2A)\mathbf{P}(X_n = 2A | A)\mathbf{P}(A)}{\mathbf{P}(X_{n+1} = 1A | A)\mathbf{P}(A)} \\ &\approx \frac{p_{2A,1A}\pi_{2A}}{\pi_{1A}} \\ &= \frac{\frac{1}{5} \cdot \frac{3}{4}}{\frac{1}{4}} \\ &= \frac{3}{5} . \end{aligned}$$

Note that the right hand side above equals $p_{1A,2A}$, as clear from the local balance equation $\pi_{1A}p_{1A,2A} = \pi_{2A}p_{2A,1A}$.

G1[†]. With $a > 0$ and $c \geq 0$,

$$\begin{aligned}
 \mathbf{P}(X - \mu \geq a) &= \mathbf{P}(X - \mu + c \geq a + c) \\
 &\leq \mathbf{P}((X - \mu + c)^2 \geq (a + c)^2) \\
 &\leq \frac{\mathbf{E}((X - \mu + c)^2)}{(a + c)^2} \\
 &= \frac{(\sigma^2 + c^2)}{(a + c)^2}
 \end{aligned}$$

where the first inequality follows from the fact that $a + c > 0$, and the second inequality follows from the Markov inequality.

To tighten the bound, we treat $(\sigma^2 + c^2)/(a + c)^2$ as a function of c , and find c such that the derivative is 0. The minimum occurs at $c = \sigma^2/a$. Therefore,

$$\mathbf{P}(X - \mu \geq a) \leq \frac{(\sigma^2 + \frac{(\sigma^4)}{a^2})}{(a + \frac{\sigma^2}{a})^2} = \frac{\sigma^2}{(\sigma^2 + a^2)}$$

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

LECTURE 19

Limit theorems – I

- **Readings:** Sections 5.1-5.3; start Section 5.4

- X_1, \dots, X_n i.i.d.

$$M_n = \frac{X_1 + \dots + X_n}{n}$$

What happens as $n \rightarrow \infty$?

- Why bother?
- A tool: Chebyshev's inequality
- Convergence "in probability"
- Convergence of M_n
(weak law of large numbers)

Chebyshev's inequality

- Random variable X
(with finite mean μ and variance σ^2)

$$\begin{aligned} \sigma^2 &= \int (x - \mu)^2 f_X(x) dx \\ &\geq \int_{-\infty}^{-c} (x - \mu)^2 f_X(x) dx + \int_c^{\infty} (x - \mu)^2 f_X(x) dx \\ &\geq c^2 \cdot \mathbf{P}(|X - \mu| \geq c) \end{aligned}$$

$$\mathbf{P}(|X - \mu| \geq c) \leq \frac{\sigma^2}{c^2}$$

$$\mathbf{P}(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

Deterministic limits

- Sequence a_n
Number a
- a_n converges to a

$$\lim_{n \rightarrow \infty} a_n = a$$

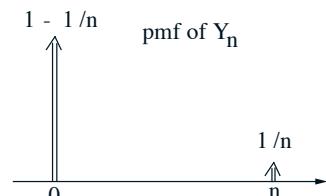
" a_n eventually gets and stays (arbitrarily) close to a "

- For every $\epsilon > 0$, there exists n_0 , such that for every $n \geq n_0$, we have $|a_n - a| \leq \epsilon$.

Convergence "in probability"

- Sequence of random variables Y_n
- converges in probability to a number a : "(almost all) of the PMF/PDF of Y_n , eventually gets concentrated (arbitrarily) close to a "
- For every $\epsilon > 0$,

$$\lim_{n \rightarrow \infty} \mathbf{P}(|Y_n - a| \geq \epsilon) = 0$$



Does Y_n converge?

Convergence of the sample mean

(Weak law of large numbers)

- X_1, X_2, \dots i.i.d.
finite mean μ and variance σ^2

$$M_n = \frac{X_1 + \cdots + X_n}{n}$$

- $E[M_n] =$

- $\text{Var}(M_n) =$

$$P(|M_n - \mu| \geq \epsilon) \leq \frac{\text{Var}(M_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}$$

- M_n converges in probability to μ

The pollster's problem

- f : fraction of population that “...”
- i th (randomly selected) person polled:

$$X_i = \begin{cases} 1, & \text{if yes,} \\ 0, & \text{if no.} \end{cases}$$

- $M_n = (X_1 + \cdots + X_n)/n$
fraction of “yes” in our sample
- Goal: 95% confidence of $\leq 1\%$ error

$$P(|M_n - f| \geq .01) \leq .05$$

- Use Chebyshev's inequality:

$$\begin{aligned} P(|M_n - f| \geq .01) &\leq \frac{\sigma_{M_n}^2}{(0.01)^2} \\ &= \frac{\sigma_x^2}{n(0.01)^2} \leq \frac{1}{4n(0.01)^2} \end{aligned}$$

- If $n = 50,000$,
then $P(|M_n - f| \geq .01) \leq .05$
(conservative)

Different scalings of M_n

- X_1, \dots, X_n i.i.d.
finite variance σ^2
- Look at three variants of their sum:
- $S_n = X_1 + \cdots + X_n$ variance $n\sigma^2$
- $M_n = \frac{S_n}{n}$ variance σ^2/n
converges “in probability” to $E[X]$ (WLLN)
- $\frac{S_n}{\sqrt{n}}$ constant variance σ^2
– Asymptotic shape?

The central limit theorem

- “Standardized” $S_n = X_1 + \cdots + X_n$:

$$Z_n = \frac{S_n - E[S_n]}{\sigma_{S_n}} = \frac{S_n - nE[X]}{\sqrt{n}\sigma}$$

- zero mean
- unit variance

- Let Z be a standard normal r.v.
(zero mean, unit variance)

- **Theorem:** For every c :

$$P(Z_n \leq c) \rightarrow P(Z \leq c)$$

- $P(Z \leq c)$ is the standard normal CDF,
 $\Phi(c)$, available from the normal tables

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013

Transcript – Lecture 20

The following content is provided under a Creative Commons license. Your support will help MIT OpenCourseWare continue to offer high-quality educational resources for free. To make a donation or view additional materials from hundreds of MIT courses, visit MIT OpenCourseWare at ocw.mit.edu.

PROFESSOR: We're going to finish today our discussion of limit theorems. I'm going to remind you what the central limit theorem is, which we introduced briefly last time. We're going to discuss what exactly it says and its implications. And then we're going to apply to a couple of examples, mostly on the binomial distribution.

OK, so the situation is that we are dealing with a large number of independent, identically distributed random variables. And we want to look at the sum of them and say something about the distribution of the sum. We might want to say that the sum is distributed approximately as a normal random variable, although, formally, this is not quite right. As n goes to infinity, the distribution of the sum becomes very spread out, and it doesn't converge to a limiting distribution.

In order to get an interesting limit, we need first to take the sum and standardize it. By standardizing it, what we mean is to subtract the mean and then divide by the standard deviation. Now, the mean is, of course, n times the expected value of each one of the X 's. And the standard deviation is the square root of the variance. The variance is n times sigma squared, where sigma is the variance of the X 's -- so that's the standard deviation.

And after we do this, we obtain a random variable that has 0 mean -- its centered -- and the variance is equal to 1. And so the variance stays the same, no matter how large n is going to be.

So the distribution of Z_n keeps changing with n , but it cannot change too much. It stays in place. The mean is 0, and the width remains also roughly the same because the variance is 1. The surprising thing is that, as n grows, that distribution of Z_n kind of settles in a certain asymptotic shape. And that's the shape of a standard normal random variable. So standard normal means that it has 0 mean and unit variance.

More precisely, what the central limit theorem tells us is a relation between the cumulative distribution function of Z_n and its relation to the cumulative distribution function of the standard normal. So for any given number, c , the probability that Z_n is less than or equal to c , in the limit, becomes the same as the probability that the standard normal becomes less than or equal to c . And of course, this is useful because these probabilities are available from the normal tables, whereas the distribution of Z_n might be a very complicated expression if you were to calculate it exactly.

So some comments about the central limit theorem. First thing is that it's quite amazing that it's universal. It doesn't matter what the distribution of the X 's is. It can be any distribution

whatsoever, as long as it has finite mean and finite variance. And when you go and do your approximations using the central limit theorem, the only thing that you need to know about the distribution of the X 's are the mean and the variance. You need those in order to standardize S_n . I mean -- to subtract the mean and divide by the standard deviation -- you need to know the mean and the variance. But these are the only things that you need to know in order to apply it.

In addition, it's a very accurate computational shortcut. So the distribution of this Z_n 's, in principle, you can calculate it by convolution of the distribution of the X 's with itself many, many times. But this is tedious, and if you try to do it analytically, it might be a very complicated expression. Whereas by just appealing to the standard normal table for the standard normal random variable, things are done in a very quick way. So it's a nice computational shortcut if you don't want to get an exact answer to a probability problem.

Now, at a more philosophical level, it justifies why we are really interested in normal random variables. Whenever you have a phenomenon which is noisy, and the noise that you observe is created by adding the lots of little pieces of randomness that are independent of each other, the overall effect that you're going to observe can be described by a normal random variable. So in a classic example that goes 100 years back or so, suppose that you have a fluid, and inside that fluid, there's a little particle of dust or whatever that's suspended in there. That little particle gets hit by molecules completely at random -- and so what you're going to see is that particle kind of moving randomly inside that liquid.

Now that random motion, if you ask, after one second, how much is my particle displaced, let's say, in the x -axis along the x direction. That displacement is very, very well modeled by a normal random variable. And the reason is that the position of that particle is decided by the cumulative effect of lots of random hits by molecules that hit that particle.

So that's a sort of celebrated physical model that goes under the name of Brownian motion. And it's the same model that some people use to describe the movement in the financial markets. The argument might go that the movement of prices has to do with lots of little decisions and lots of little events by many, many different actors that are involved in the market. So the distribution of stock prices might be well described by normal random variables. At least that's what people wanted to believe until somewhat recently.

Now, the evidence is that, actually, these distributions are a little more heavy-tailed in the sense that extreme events are a little more likely to occur than what normal random variables would seem to indicate. But as a first model, again, it could be a plausible argument to have, at least as a starting model, one that involves normal random variables. So this is the philosophical side of things.

On the more accurate, mathematical side, it's important to appreciate exactly what kind of statement the central limit theorem is. It's a statement about the convergence of the CDF of these standardized random variables to the CDF of a normal. So it's a statement about convergence of CDFs. It's not a statement about convergence of PMFs, or convergence of PDFs.

Now, if one makes additional mathematical assumptions, there are variations of the central limit theorem that talk about PDFs and PMFs. But in general, that's not necessarily the case. And I'm going to illustrate this with-- I have a plot here which is not in your slides. But just to make the point, consider two different discrete distributions.

This discrete distribution takes values 1, 4, 7. This discrete distribution can take values 1, 2, 4, 6, and 7. So this one has sort of a periodicity of 3, this one, the range of values is a little more interesting. The numbers in these two distributions are cooked up so that they have the same mean and the same variance.

Now, what I'm going to do is to take eight independent copies of the random variable and plot the PMF of the sum of eight random variables. Now, if I plot the PMF of the sum of 8 of these, I get the plot, which corresponds to these bullets in this diagram. If I take 8 random variables, according to this distribution, and add them up and compute their PMF, the PMF I get is the one denoted here by the X's. The two PMFs look really different, at least, when you eyeball them.

On the other hand, if you were to plot the CDFs of them, then the CDFs, if you compare them with the normal CDF, which is this continuous curve, the CDF, of course, it goes up in steps because we're looking at discrete random variables. But it's very close to the normal CDF. And if we, instead of n equal to 8, we were to take 16, then the coincidence would be even better.

So in terms of CDFs, when we add 8 or 16 of these, we get very close to the normal CDF. We would get essentially the same picture if I were to take 8 or 16 of these. So the CDFs sit, essentially, on top of each other, although the two PMFs look quite different. So this is to appreciate that, formally speaking, we only have a statement about CDFs, not about PMFs.

Now in practice, how do you use the central limit theorem? Well, it tells us that we can calculate probabilities by treating Z_n as if it were a standard normal random variable. Now Z_n is a linear function of S_n . Conversely, S_n is a linear function of Z_n . Linear functions of normals are normal. So if I pretend that Z_n is normal, it's essentially the same as if we pretend that S_n is normal. And so we can calculate probabilities that have to do with S_n as if S_n were normal. Now, the central limit theorem does not tell us that S_n is approximately normal. The formal statement is about Z_n , but, practically speaking, when you use the result, you can just pretend that S_n is normal.

Finally, it's a limit theorem, so it tells us about what happens when n goes to infinity. If we are to use it in practice, of course, n is not going to be infinity. Maybe n is equal to 15. Can we use a limit theorem when n is a small number, as small as 15?

Well, it turns out that it's a very good approximation. Even for quite small values of n , it gives us very accurate answers. So n over the order of 15, or 20, or so give us very good results in practice. There are no good theorems that will give us hard guarantees because the quality of the approximation does depend on the details of the distribution of the X's. If the X's have a distribution that, from the outset, looks a little bit like the normal, then for small values of n , you are going to see, essentially, a normal distribution for the sum. If the distribution of the X's is very different from the normal, it's going to take a larger value of n for the central limit theorem to take effect.

So let's illustrate this with a few representative plots. So here, we're starting with a discrete uniform distribution that goes from 1 to 8. Let's add 2 of these random variables, 2 random variables with this PMF, and find the PMF of the sum. This is a convolution of 2 discrete uniforms, and I believe you have seen this exercise before. When you convolve this with itself, you get a triangle. So this is the PMF for the sum of two discrete uniforms.

Now let's continue. Let's convolve this with itself. These was going to give us the PMF of a sum of 4 discrete uniforms. And we get this, which starts looking like a normal. If we go to n equal to 32, then it looks, essentially, exactly like a normal. And it's an excellent approximation. So this is the PMF of the sum of 32 discrete random variables with this uniform distribution.

If we start with a PMF which is not symmetric-- this one is symmetric around the mean. But if we start with a PMF which is non-symmetric, so this is, here, is a truncated geometric PMF, then things do not work out as nicely when I add 8 of these. That is, if I convolve this with itself 8 times, I get this PMF, which maybe resembles a little bit to the normal one.

But you can really tell that it's different from the normal if you focus at the details here and there. Here it sort of rises sharply. Here it tails off a bit slower. So there's an asymmetry here that's present, and which is a consequence of the asymmetry of the distribution we started with. If we go to 16, it looks a little better, but still you can see the asymmetry between this tail and that tail.

If you get to 32 there's still a little bit of asymmetry, but at least now it starts looking like a normal distribution. So the moral from these plots is that it might vary, a little bit, what kind of values of n you need before you get the really good approximation. But for values of n in the range 20 to 30 or so, usually you expect to get a pretty good approximation. At least that's what the visual inspection of these graphs tells us.

So now that we know that we have a good approximation in our hands, let's use it. Let's use it by revisiting an example from last time. This is the polling problem. We're interested in the fraction of population that has a certain habit been. And we try to find what f is. And the way we do it is by polling people at random and recording the answers that they give, whether they have the habit or not. So for each person, we get the Bernoulli random variable. With probability f , a person is going to respond 1, or yes, so this is with probability f . And with the remaining probability $1-f$, the person responds no.

We record this number, which is how many people answered yes, divided by the total number of people. That's the fraction of the population that we asked. This is the fraction inside our sample that answered yes. And as we discussed last time, you might start with some specs for the poll. And the specs have two parameters-- the accuracy that you want and the confidence that you want to have that you did really obtain the desired accuracy. So the specs here is that we want, probability 95% that our estimate is within 1 % point from the true answer.

So the event of interest is this. That's the result of the poll minus distance from the true answer is less or bigger than 1 % point. And we're interested in calculating or approximating this particular probability.

So we want to do it using the central limit theorem. And one way of arranging the mechanics of this calculation is to take the event of interest and massage it by subtracting and dividing things from both sides of this inequality so that you bring him to the picture the standardized random variable, the Z_n , and then apply the central limit theorem.

So the event of interest, let me write it in full, M_n is this quantity, so I'm putting it here, minus f , which is the same as nf divided by n . So this is the same as that event. We're going to calculate the probability of this. This is not exactly in the form in which we apply the central limit theorem. To apply the central limit theorem, we need, down here, to have $\sigma \sqrt{n}$.

So how can I put $\sigma \sqrt{n}$ here? I can divide both sides of this inequality by σ . And then I can take a factor of \sqrt{n} from here and send it to the other side.

So this event is the same as that event. This will happen if and only if that will happen. So calculating the probability of this event here is the same as calculating the probability that this event happens.

And now we are in business because the random variable that we got in here is Z_n , or the absolute value of Z_n , and we're talking about the probability that Z_n , absolute value of Z_n , is bigger than a certain number. Since Z_n is to be approximated by a standard normal random variable, our approximation is going to be, instead of asking for Z_n being bigger than this number, we will ask for Z , absolute value of Z , being bigger than this number.

So this is the probability that we want to calculate. And now Z is a standard normal random variable. There's a small difficulty, the one that we also encountered last time. And the difficulty is that the standard deviation, σ , of the X_i 's is not known. σ is equal to f times σ , in this example, is f times $(1-f)$, and the only thing that we know about σ is that it's going to be a number less than $1/2$.

OK, so we're going to have to use an inequality here. We're going to use a conservative value of σ , the value of σ equal to $1/2$ and use that instead of the exact value of σ . And this gives us an inequality going this way.

Let's just make sure why the inequality goes this way. We got, on our axis, two numbers. One number is $0.01\sqrt{n}/\sigma$ and the other number is $0.02\sqrt{n}/\sigma$. And my claim is that the numbers are related to each other in this particular way.

Why is this? σ is less than $1/2$. So $1/\sigma$ is bigger than 2 . So since $1/\sigma$ is bigger than 2 this means that this number sits to the right of that number. So here we have the probability that Z is bigger than this number. The probability of falling out there is less than the probability of falling in this interval.

So that's what that last inequality is saying-- this probability is smaller than that probability. This is the probability that we're interested in, but since we don't know σ , we take the conservative value, and we use an upper bound in terms of the probability of this interval here.

And now we are in business. We can start using our normal tables to calculate probabilities of interest. So for example, let's say that's we take n to be 10,000. How is the calculation going to go? We want to calculate the probability that the absolute value of Z is bigger than 0.2 times 1000, which is the probability that the absolute value of Z is larger than or equal to 2.

And here let's do some mechanics, just to stay in shape. The probability that you're larger than or equal to 2 in absolute value, since the normal is symmetric around the mean, this is going to be twice the probability that Z is larger than or equal to 2.

Can we use the cumulative distribution function of Z to calculate this? Well, almost the cumulative gives us probabilities of being less than something, not bigger than something. So we need one more step and write this as 1 minus the probability that Z is less than or equal to 2.

And this probability, now, you can read off from the normal tables. And the normal tables will tell you that this probability is 0.9772. And you do get an answer. And the answer is 0.0456. OK, so we tried 10,000. And we find that our probably of error is 4.5%, so we're doing better than the spec that we had. So this tells us that maybe we have some leeway. Maybe we can use a smaller sample size and still stay without our specs.

Let's try to find how much we can push the envelope. How much smaller can we take n ? To answer that question, we need to do this kind of calculation, essentially, going backwards. We're going to fix this number to be 0.05 and work backwards here to find-- did I do a mistake here? 10,000. So I'm missing a 0 here. Ah, but I'm taking the square root, so it's 100.

Where did the 0.02 come in from? Ah, from here. OK, all right. 0.02 times 100, that gives us 2. OK, all right. Very good, OK. So we'll have to do this calculation now backwards, figure out if this is 0.05, what kind of number we're going to need here and then here, and from this we will be able to tell what value of n do we need.

OK, so we want to find n such that the probability that Z is bigger than 0.02 square root n is 0.05. OK, so Z is a standard normal random variable. And we want the probability that we are outside this range. We want the probability of those two tails together. Those two tails together should have probability of 0.05. This means that this tail, by itself, should have probability 0.025. And this means that this probability should be 0.975.

Now, if this probability is to be 0.975, what should that number be? You go to the normal tables, and you find which is the entry that corresponds to that number. I actually brought a normal table with me. And 0.975 is down here. And it tells you that to the number that corresponds to it is 1.96.

So this tells us that this number should be equal to 1.96. And now, from here, you do the calculations. And you find that n is 9604. So with a sample of 10,000, we got probability of error 4.5%. With a slightly smaller sample size of 9,600, we can get the probability of a mistake to be 0.05, which was exactly our spec.

So these are essentially the two ways that you're going to be using the central limit theorem. Either you're given n and you try to calculate probabilities. Or you're given the probabilities, and you want to work backwards to find n itself.

So in this example, the random variable that we dealt with was, of course, a binomial random variable. The X_i 's were Bernoulli, so the sum of the X_i 's were binomial. So the central limit theorem certainly applies to the binomial distribution. To be more precise, of course, it applies to the standardized version of the binomial random variable.

So here's what we did, essentially, in the previous example. We fixed the number p , which is the probability of success in our experiments. p corresponds to f in the previous example. Let every X_i a Bernoulli random variable and are standing assumption is that these random variables are independent. When we add them, we get a random variable that has a binomial distribution. We know the mean and the variance of the binomial, so we take S_n , we subtract the mean, which is this, divide by the standard deviation. The central limit theorem tells us that the cumulative distribution function of this random variable is a standard normal random variable in the limit.

So let's do one more example of a calculation. Let's take n to be-- let's choose some specific numbers to work with. So in this example, first thing to do is to find the expected value of S_n , which is n times p . It's 18.

Then we need to write down the standard deviation. The variance of S_n is the sum of the variances. It's np times $(1-p)$. And in this particular example, p times $(1-p)$ is $1/4$, n is 36, so this is 9. And that tells us that the standard deviation of this n is equal to 3.

So what we're going to do is to take the event of interest, which is S_n less than 21, and rewrite it in a way that involves the standardized random variable. So to do that, we need to subtract the mean. So we write this as $S_n - 3$ should be less than or equal to $21 - 3$. This is the same event. And then divide by the standard deviation, which is 3, and we end up with this. So the event itself of--

AUDIENCE: [INAUDIBLE].

Should subtract, 18, yes, which gives me a much nicer number out here, which is 1. So the event of interest, that S_n is less than 21, is the same as the event that a standard normal random variable is less than or equal to 1. And once more, you can look this up at the normal tables. And you find that the answer that you get is 0.43.

Now it's interesting to compare this answer that we got through the central limit theorem with the exact answer. The exact answer involves the exact binomial distribution. What we have here is the binomial probability that, S_n is equal to k . S_n being equal to k is given by this formula. And we add, over all values for k going from 0 up to 21, we write a two lines code to calculate this sum, and we get the exact answer, which is 0.8785. So there's a pretty good agreement between the two, although you wouldn't call that's necessarily excellent agreement.

Can we do a little better than that? OK. It turns out that we can. And here's the idea. So our random variable S_n has a mean of 18. It has a binomial distribution. It's described by a PMF that has a shape roughly like this and which keeps going on.

Using the central limit theorem is basically pretending that S_n is normal with the right mean and variance. So pretending that Z_n has 0 mean unit variance, we approximate it with Z , that has 0 mean unit variance. If you were to pretend that S_n is normal, you would approximate it with a normal that has the correct mean and correct variance. So it would still be centered at 18. And it would have the same variance as the binomial PMF.

So using the central limit theorem essentially means that we keep the mean and the variance what they are but we pretend that our distribution is normal. We want to calculate the probability that S_n is less than or equal to 21. I pretend that my random variable is normal, so I draw a line here and I calculate the area under the normal curve going up to 21. That's essentially what we did.

Now, a smart person comes around and says, S_n is a discrete random variable. So the event that S_n is less than or equal to 21 is the same as S_n being strictly less than 22 because nothing in between can happen. So I'm going to use the central limit theorem approximation by pretending again that S_n is normal and finding the probability of this event while pretending that S_n is normal. So what this person would do would be to draw a line here, at 22, and calculate the area under the normal curve all the way to 22.

Who is right? Which one is better? Well neither, but we can do better than both if we sort of split the difference. So another way of writing the same event for S_n is to write it as S_n being less than 21.5. In terms of the discrete random variable S_n , all three of these are exactly the same event. But when you do the continuous approximation, they give you different probabilities. It's a matter of whether you integrate the area under the normal curve up to here, up to the midway point, or up to 22. It turns out that integrating up to the midpoint is what gives us the better numerical results. So we take here 21 and 1/2, and we integrate the area under the normal curve up to here.

So let's do this calculation and see what we get. What would we change here? Instead of 21, we would now write 21 and 1/2. This 18 becomes, no, that 18 stays what it is. But this 21 becomes 21 and 1/2. And so this one becomes $1 + 0.5$ by 3. This is 117.

So we now look up into the normal tables and ask for the probability that Z is less than 1.17. So this here gets approximated by the probability that the standard normal is less than 1.17. And the normal tables will tell us this is 0.879.

Going back to the previous slide, what we got this time with this improved approximation is 0.879. This is a really good approximation of the correct number. This is what we got using the 21. This is what we get using the 21 and 1/2. And it's an approximation that's sort of right on-- a very good one.

The moral from this numerical example is that doing this 1 and 1/2 correction does give us better approximations. In fact, we can use this 1/2 idea to even calculate individual probabilities. So suppose you want to approximate the probability that S_n equal to 19. If you were to pretend that S_n is normal and calculate this probability, the probability that the normal random variable is equal to 19 is 0. So you don't get an interesting answer.

You get a more interesting answer by writing this event, 19 as being the same as the event of falling between 18 and 1/2 and 19 and 1/2 and using the normal approximation to calculate this probability. In terms of our previous picture, this corresponds to the following.

We are interested in the probability that S_n is equal to 19. So we're interested in the height of this bar. We're going to consider the area under the normal curve going from here to here, and use this area as an approximation for the height of that particular bar.

So what we're basically doing is, we take the probability under the normal curve that's assigned over a continuum of values and attributed it to different discrete values. Whatever is above the midpoint gets attributed to 19. Whatever is below that midpoint gets attributed to 18. So this is green area is our approximation of the value of the PMF at 19.

So similarly, if you wanted to approximate the value of the PMF at this point, you would take this interval and integrate the area under the normal curve over that interval. It turns out that this gives a very good approximation of the PMF of the binomial. And actually, this was the context in which the central limit theorem was proved in the first place, when this business started.

So this business goes back a few hundred years. And the central limit theorem was first approved by considering the PMF of a binomial random variable when p is equal to 1/2. People did the algebra, and they found out that the exact expression for the PMF is quite well approximated by that expression hat you would get from a normal distribution. Then the proof was extended to binomials for more general values of p .

So here we talk about this as a refinement of the general central limit theorem, but, historically, that refinement was where the whole business got started in the first place. All right, so let's go through the mechanics of approximating the probability that S_n is equal to 19-- exactly 19. As we said, we're going to write this event as an event that covers an interval of unit length from 18 and 1/2 to 19 and 1/2. This is the event of interest.

First step is to massage the event of interest so that it involves our Z_n random variable. So subtract 18 from all sides. Divide by the standard deviation of 3 from all sides. That's the equivalent representation of the event. This is our standardized random variable Z_n . These are just these numbers.

And to do an approximation, we want to find the probability of this event, but Z_n is approximately normal, so we plug in here the Z , which is the standard normal. So we want to find the probability that the standard normal falls inside this interval. You find these using CDFs because this is the probability that you're less than this but not less than that. So it's a difference between two cumulative probabilities.

Then, you look up your normal tables. You find two numbers for these quantities, and, finally, you get a numerical answer for an individual entry of the PMF of the binomial. This is a pretty good approximation, it turns out. If you were to do the calculations using the exact formula, you would get something which is pretty close-- an error in the third digit-- this is pretty good.

So I guess what we did here with our discussion of the binomial slightly contradicts what I said before-- that the central limit theorem is a statement about cumulative distribution functions. In general, it doesn't tell you what to do to approximate PMFs themselves. And that's indeed the case in general. One the other hand, for the special case of a binomial distribution, the central limit theorem approximation, with this $1/2$ correction, is a very good approximation even for the individual PMF.

All right, so we spent quite a bit of time on mechanics. So let's spend the last few minutes today thinking a bit and look at a small puzzle. So the puzzle is the following. Consider Poisson process that runs over a unit interval. And where the arrival rate is equal to 1. So this is the unit interval. And let X be the number of arrivals. And this is Poisson, with mean 1.

Now, let me take this interval and divide it into n little pieces. So each piece has length $1/n$. And let X_i be the number of arrivals during the i th little interval.

OK, what do we know about the random variables X_i ? Is they are themselves Poisson. It's a number of arrivals during a small interval. We also know that when n is big, so the length of the interval is small, these X_i 's are approximately Bernoulli, with mean $1/n$.

Guess it doesn't matter whether we model them as Bernoulli or not. What matters is that the X_i 's are independent. Why are they independent? Because, in a Poisson process, these joint intervals are independent of each other.

So the X_i 's are independent. And they also have the same distribution. And we have that X , the total number of arrivals, is the sum over the X_i 's. So the central limit theorem tells us that, approximately, the sum of independent, identically distributed random variables, when we have lots of these random variables, behaves like a normal random variable. So by using this decomposition of X into a sum of i.i.d random variables, and by using values of n that are bigger and bigger, by taking the limit, it should follow that X has a normal distribution.

On the other hand, we know that X has a Poisson distribution. So something must be wrong in this argument here. Can we really use the central limit theorem in this situation?

So what do we need for the central limit theorem? We need to have independent, identically distributed random variables. We have it here. We want them to have a finite mean and finite variance. We also have it here, means variances are finite.

What is another assumption that was never made explicit, but essentially was there? Or in other words, what is the flaw in this argument that uses the central limit theorem here? Any thoughts?

So in the central limit theorem, we said, consider-- fix a probability distribution, and let the X_i 's be distributed according to that probability distribution, and add a larger and larger number or X_i 's. But the underlying, unstated assumption is that we fix the distribution of the X_i 's. As we let n increase, the statistics of each X_i do not change.

Whereas here, I'm playing a trick on you. As I'm taking more and more random variables, I'm actually changing what those random variables are. When I take a larger n , the X_i 's are random variables with a different mean and different variance. So I'm adding more of these, but at the same time, in this example, I'm changing their distributions.

That's something that doesn't fit the setting of the central limit theorem. In the central limit theorem, you first fix the distribution of the X 's. You keep it fixed, and then you consider adding more and more according to that particular fixed distribution. So that's the catch. That's why the central limit theorem does not apply to this situation. And we're lucky that it doesn't apply because, otherwise, we would have a huge contradiction destroying probability theory.

OK, but now that's still leaves us with a little bit of a dilemma. Suppose that, here, essentially we're adding independent Bernoulli random variables. So the issue is that the central limit theorem has to do with asymptotics as n goes to infinity. And if we consider a binomial, and somebody gives us specific numbers about the parameters of that binomial, it might not necessarily be obvious what kind of approximation do we use.

In particular, we do have two different approximations for the binomial. If we fix p , then the binomial is the sum of Bernoulli's that come from a fixed distribution, we consider more and more of these. When we add them, the central limit theorem tells us that we get the normal distribution.

There's another sort of limit, which has the flavor of this example, in which we still deal with a binomial, sum of n Bernoulli's. We let that sum, the number of the Bernoulli's go to infinity. But each Bernoulli has a probability of success that goes to 0, and we do this in a way so that np , the expected number of successes, stays finite.

This is the situation that we dealt with when we first defined our Poisson process. We have a very, very large number so lots, of time slots, but during each time slot, there's a tiny probability of obtaining an arrival. Under that setting, in discrete time, we have a binomial distribution, or Bernoulli process, but when we take the limit, we obtain the Poisson process and the Poisson approximation.

So these are two equally valid approximations of the binomial. But they're valid in different asymptotic regimes. In one regime, we fixed p , let n go to infinity. In the other regime, we let both n and p change simultaneously.

Now, in real life, you're never dealing with the limiting situations. You're dealing with actual numbers. So if somebody tells you that the numbers are like this, then you should probably say that this is the situation that fits the Poisson description-- large number of slots with each slot having a tiny probability of success.

On the other hand, if p is something like this, and n is 500, then you expect to get the distribution for the number of successes. It's going to have a mean of 50 and to have a fair amount of spread around there. It turns out that the normal approximation would be better in this context.

As a rule of thumb, if n times p is bigger than 10 or 20, you can start using the normal approximation. If n times p is a small number, then you prefer to use the Poisson approximation. But there's no hard theorems or rules about how to go about this. OK, so from next time we're going to switch base again. And we're going to put together everything we learned in this class to start solving inference problems.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013 Transcript – Recitation: Probability Bounds

In this problem, we're given a collection of 10 variables, x_1 through x_{10} , where each i , x_i , is a uniform random variable between 0 and 1. So each i is uniform between 0 and 1, and all 10 variables are independent. And we'd like to develop a bound on the probability that some of the 10 variables, 1 to 10, being greater than 7 using different methods. So in part A we'll be using the Markov's inequality written here. That is, if we have a random variable, positive random variable x , the probability x is greater than a , where a is again some positive number, is bounded above by the expected value of x divided by a .

And let's see how that works out in our situation. In our situation, we will call x the summation of i equal to 1 to 10 x_i , and therefore, E of x is simply 10 times E of x_1 , the individual ones, and this gives us 5. Here we used the linearity of expectation such that the expectation of the sum of the random variable is simply the sum of the expectations. Now, we can invoke Markov's Inequality. It says x greater or equal to 7. This is less than E of x over 7, and this gives us 5 over 7.

For part B, let's see if we can improve the bound we got in part A using the Chebyshev inequality, which takes into account the variance of random variable x . Again, to refresh you on this, the Chebyshev Inequality says the probability that x deviates from its mean E of x , by more than a is bound above by the variance of x divided by a squared.

So we have to actually do some work to transform the probability we're interested in, which is x greater or equal to 7, into the form that's convenient to use using the Chebyshev Inequality. To do so, we'll rewrite this probability as the probability of x minus 5 greater or equal to 2 simply by moving 5 from the right to the left. The reason we chose 5 is because 5 is equal to the expected value of x from part A as we know before.

And in fact, this quantity is also equal to the probability that x minus 5 less or equal to negative 2. To see why this is true, recall that x is simply the summation of the x_i 's, the 10 x_i 's, and each x_i is a uniform random variable between 0 and 1. And therefore, each x_i , the distribution of which is symmetric around its mean 1/2. So we can see that after we add up all the x_i 's, the resulting distribution x is also symmetric around its mean 5. And as a result, the probability of x minus 5 greater than 2 is now equal to the probability that x minus 5 less than negative 2.

And knowing these two, we can then say they're both equal to 1/2 the probability x minus 5 absolute value greater or equal to 2, because this term right here is simply the sum of both terms here and here. At this point, we have transformed the probability of x greater or equal to 7 into the form right here, such that we can apply the Chebyshev's Inequality basically directly.

And we'll write the probably here being less than or equal to 1/2 times, applying the Chebyshev Inequality, variance of x divided by 2 squared. Now, 2 is the same as a right here, and this gives us 1/8 times-- now, the variance of x , we know is 10 times the variance of a uniform random variable between 0 and 1, which is 1/12, and that gives us 5/48.

Now, let's compare this with the number we got earlier using the Markov Inequality, which was $5/7$. We see that $5/48$ is much smaller, and this tells us that, at least for this example, using the Chebyshev Inequality combined with the information of the variance of x , we're able to get a stronger upper bound on the probability of the event that we're interested in.

Now, in part B, we saw that by using the additional information of the variance combined with the Chebyshev Inequality, we can improve upon bound given by Markov's Inequality. Now, in part C, we'll use a somewhat more powerful approach in addition to the Chebyshev Inequality, the so-called central limit theorem. Let's see if we can even get a better bound.

To remind you what a central limit theorem is, let's say we have a summation of i equal to 1 to some number n of independent and identically distributed random variables x_i . Now, the central limit theorem says the following. We take the sum right here, and subtract out its means, which is E of the same summation, and further, we'll divide out, what we call normalize, by the standard deviation of the summation. In other words, the square root of the variance of the sum of x_i .

So if we perform this procedure right here, then as the number of terms in the sums going to infinity, here as in goes to infinity, we will actually see that this random variable will converge in distribution in some way that will eventually look like a standard normal random variable with means 0 and 1. And since we know how the distribution of a standard normal looks like, we can go to table and look up certain properties of the resulting distribution. So that is a plan to do. So right now, we have about 10 variables. It's not that many compared to a huge numbering, but again, if we believe it's a good approximation, we can get some information out of it by using the central limit theorem.

So we are interesting knowing that probability summation of i equal to 1 to 10 x_1 greater or equal to 7. We'll rewrite this as 1 minus the probability the summation i equal to 1 to 10, and x_i less equal to 7. Now, we're going to apply the scaling to the summation right here. So this is equal to 1 minus the probability summation i equal to 1 to 10 x_i minus 5. Because we know from previous parts that 5 is the expected value of the sum right here, and divided by square root of 10/12.

Again, earlier we know that 10/12 is the variance of the sum of x_i 's. And we'll do the same on the other side, writing it 7 minus 5 divided by square root of 10/12. Now, if we compute out the quantity right here, we know that this quantity is roughly 2.19, and by the central limit theorem, if we believe 10 is a large enough number, then this will be roughly equal to 1 minus the CDF of a standard normal evaluated at 2.19. And we could look up the number in the table, and this gives us number roughly, 0.014.

Now let's do a quick summary of what this problem is about. We're asked to compute the probability of x greater or equal to 7, where x is the sum of 10 uniform random variables between 0 and 1, so we'll call it x_i . We know that because each random variable has expectation 1/2, adding 10 of them up, gives us expectation of 5. So this is essentially asking, what is the chance that x is more than two away from its expectation?

So if this is a real line, and 5 is here, maybe x has some distribution around 5, so the center what the expected value is at 5, we wonder how likely is it for us to see something greater than 7? Now, let's see where do we land on the probably spectrum from 0 to 1. Well, without using any information, we know the probability cannot be greater than 1, so a trivial upper bound for the probability right here will be 1.

Well, for the first part we use Markov's Inequality and that gives us some number, which is roughly equal to 0.7. In fact, we got number $5/7$, and this is from Markov's Inequality. Oh, it's better than 1, already telling us it cannot be between 0.7 and 1, but can we do better?

Well, the part B, we see that all the way, using the additional information variance, we can get this number down to $5/48$, which is roughly 0.1. Already, that's much better than 0.7. Can we even do better? And this is the Chebyshev, and it turns out we can indeed do better. Using the central limit theorem, we can squeeze this number all the way down to 0.014, almost a 10 times improvement over the previous number. This is from central limit theorem.

As we can see, by using different bounding techniques, we can progressively improve the bound on the probability of x exceeding 7, and from this problem we learned that even with 10 variables, the truth is more like this, which says that the distribution of x concentrates very heavily around 5, and hence, the probability of x being greater or equal to 7 could be much smaller than one might expect.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013 Transcript – Recitation: Using the Central Limit Theorem

Hi. In this video, we're going to do some approximate calculations using the central limit theorem.

We're given that X_n is the number of gadgets produced on day n by a factory. And it has a normal distribution with mean 5 and variance 9. And they're all independent and identically distributed.

We're looking for the probability that the total number of gadgets in 100 days is less than 440.

To start, we can first write this as the probability of the sum of the gadgets produced on each of 100 days being less than 440. Notice that this is a sum of a large number of independent random variables. So we can use the central limit theorem and approximate the sum as a normal random variable.

And then, basically, in order to compute this probability, we'd basically need to standardize this and then use the standard normal table.

So let's first compute the expectation and variance of the sum. So I'm going to actually sum up from 1 to n instead of 100, to do it more generally. So the linearity is preserved for the expectation operator. So this is the sum of the expected value. And since they're all identically distributed, they all have the same expectation, and there are n of them. And so we have this being n times 5.

For the variance of the sum is also the sum of the variances because the independents. And so they're identically distributed to the -- so we have n times the variance of X_i , and this is n times 9.

So now, we can standardize it, or make it 0 mean and variance 1. So to do that we would take these X_i 's, subtract by their mean. So it's going to be 5 times 100 of them, so it's 500 over the square root of the variance, which is going to be 9 times 100 of them, so it's going to be 900. So that's going to be less than 440 minus 500 over square root of 900.

So notice what we're trying to do here is-- notice that the sum of X_i 's is a discrete quantity. So it's a discrete random variable, so it may have a PMF like this. And we're trying to approximate it with a normal density. So this is not drawn to scale, but let's say that this is 440 and this is 439. Basically, we're trying to say what's the probability of this being less than 440, so it's the probability that it's 439, or 438, or 437.

But in the continuous case, a good approximation to this would be to take the middle, say, 439.5, and compute the area below that.

So in this case, when we do the normal approximation, it works out better if we use this half correction. And so, this, in this case, probability, let's call Z the standard normal. And so this is approximately equal to a standard normal with the probability of standard normal being less than whatever that is. And if you plug that into your calculator, you get negative 2.02.

So now, if we try to figure out what this-- from the table, we'll find that negative values are not tabulated. But we know that the normal, the center of normal is symmetric, and so if we want to compute the area in this region, it's the same as the area in this region, above 2.02.

So this is the same as the probability that Z is bigger than 2.02. That's just 1 minus the probability that Z is less than or equal to 2.02, and so that's, by definition, phi of 2.02. And if we look it up on the table, 2.02 has probability here of 0.9783. And we can just write that in. That's the answer for Part A.

So now for Part B.

We're asked what's the largest n, approximately, so that it satisfies this.

So again, we can use the central limit theorem. Use similar steps here so that we have, in this case, n greater than or equal to 200 plus 5n. And standardized. So we have n and the mean here-- this is where this comes handy. It's going to be 5n and the variance is 9n. It's greater than or equal to. 5n's will cancel and you subtract. And then you get 200 over the square root of 9n.

And we can, again, use the half approximation here, half correction here. But I'm not going to do it, to keep the problem simple. And so in this case, this is approximately equal to the standard normal being greater than probability of the center of normal being greater than or equal to 200 over square root of 9n.

And so same sort of thing here. This is just 1 minus this. The equal sign doesn't matter because Z is a continuous random variable. And so we have this here.

And we want this to be less than or equal 0.05. So that means that phi of 200 over square root of 9 has to be greater than or equal to 0.95. So we're basically looking for something here that ensures that this region's at least 0.95.

So if you look at the table, 0.95 lies somewhere in between 1.64 and 1.65. And I'm going to use 1.65 to be conservative, because we want this region to be at least 0.95. So 1.65 works better here.

And so we want this thing, this here, which is going to be 200 over square root of n-- square root of 9n, to be bigger than or equal to 1.65. So n here is going to be less than or equal to 200 over 1.65 squared, 1 over 9.

If you plug this into your calculator, you might have a decimal in there. Then we just pick n, the largest integer that satisfies this. So we can plug that into your calculator, you'll find that it's going to be 1,632. That's part B.

Last part.

Let n be the first day when the total number of gadgets is greater than 1,000. What's the probability that n is greater than or equal to 220?

Again, we want to use the central limit theorem, but the trick here is to recognize that this is actually equal to the probability that the sum from i equals 1 to 219 of X_i , is less than or equal to 1,000.

So let's look at both directions to check this.

If n is greater than or equal to 220, then this has to be true. Because if it weren't true, and if this were greater than 1,000, then n would have been less than or equal to 219. So this direction works.

The other direction. If this were the case, it has to be the case that n is greater than or equal to 220, because up till 219 it hasn't exceeded 1,000. And so, at some point beyond that, it's going to exceed 1,000 and n is going to be greater than or equal to 220. So this is the key trick here.

And once you see this, you realize that this is very easy because we do the same steps as we did before. So you're looking for this, this is equal to, again, you do your standardization. So this is from 219, and you get 5 times 219 for the mean, and 9 times 219 for the variance, less than or equal to 1,000 minus 5 times 219 over square root of 9 times 219.

Again, you can do the half correction here, make it 1,000.5, but I'm not going to do that in this case, for simplicity. So this is approximately equal to Z being less than whatever this is. And if you plug it in, you'll find that this is negative 2.14.

So in this case, we have this is the probability that Z -- again, we do the same thing-- is greater than or equal to 2.14. And this is 1 minus the probability that Z is less than or equal to 2.14. And that's just ϕ of 2.14-- 1 minus ϕ of 2.14. And that's-- if you look it up on the table, it's 2.14. It's 0.9838. So here's your answer.

So we're done with Part C as well. So in this exercise, we did a lot of approximate calculations using the central--

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

LECTURE 20

THE CENTRAL LIMIT THEOREM

- Readings: Section 5.4
- X_1, \dots, X_n i.i.d., finite variance σ^2

- “Standardized” $S_n = X_1 + \dots + X_n$:

$$Z_n = \frac{S_n - \mathbb{E}[S_n]}{\sigma_{S_n}} = \frac{S_n - n\mathbb{E}[X]}{\sqrt{n}\sigma}$$

- $\mathbb{E}[Z_n] = 0$, $\text{var}(Z_n) = 1$

- Let Z be a standard normal r.v. (zero mean, unit variance)

- **Theorem:** For every c :

$$\mathbf{P}(Z_n \leq c) \rightarrow \mathbf{P}(Z \leq c)$$

- $\mathbf{P}(Z \leq c)$ is the standard normal CDF, $\Phi(c)$, available from the normal tables

Usefulness

- universal; only means, variances matter
- accurate computational shortcut
- justification of normal models

What exactly does it say?

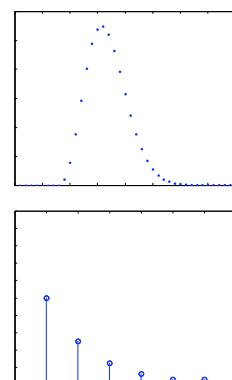
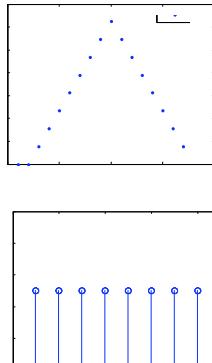
- CDF of Z_n converges to normal CDF
 - not a statement about convergence of PDFs or PMFs

Normal approximation

- Treat Z_n as if normal
 - also treat S_n as if normal

Can we use it when n is “moderate”?

- Yes, but no nice theorems to this effect
- Symmetry helps a lot



The pollster's problem using the CLT

- f : fraction of population that “...”
- i th (randomly selected) person polled:

$$X_i = \begin{cases} 1, & \text{if yes,} \\ 0, & \text{if no.} \end{cases}$$

- $M_n = (X_1 + \dots + X_n)/n$

- Suppose we want:

$$\mathbf{P}(|M_n - f| \geq .01) \leq .05$$

- Event of interest: $|M_n - f| \geq .01$

$$\left| \frac{X_1 + \dots + X_n - nf}{n} \right| \geq .01$$

$$\left| \frac{X_1 + \dots + X_n - nf}{\sqrt{n}\sigma} \right| \geq \frac{.01\sqrt{n}}{\sigma}$$

$$\begin{aligned} \mathbf{P}(|M_n - f| \geq .01) &\approx \mathbf{P}(|Z| \geq .01\sqrt{n}/\sigma) \\ &\leq \mathbf{P}(|Z| \geq .02\sqrt{n}) \end{aligned}$$

Apply to binomial

- Fix p , where $0 < p < 1$
- X_i : Bernoulli(p)
- $S_n = X_1 + \dots + X_n$: Binomial(n, p)
 - mean np , variance $np(1 - p)$
- CDF of $\frac{S_n - np}{\sqrt{np(1 - p)}}$ → standard normal

Example

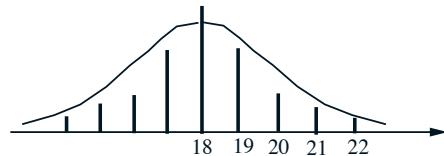
- $n = 36, p = 0.5$; find $P(S_n \leq 21)$

- Exact answer:

$$\sum_{k=0}^{21} \binom{36}{k} \left(\frac{1}{2}\right)^{36} = 0.8785$$

The 1/2 correction for binomial approximation

- $P(S_n \leq 21) = P(S_n < 22)$, because S_n is integer
- Compromise: consider $P(S_n \leq 21.5)$



De Moivre–Laplace CLT (for binomial)

- When the 1/2 correction is used, CLT can also approximate the binomial p.m.f. (not just the binomial CDF)

$$P(S_n = 19) = P(18.5 \leq S_n \leq 19.5)$$

$$18.5 \leq S_n \leq 19.5 \iff$$

$$\frac{18.5 - 18}{3} \leq \frac{S_n - 18}{3} \leq \frac{19.5 - 18}{3} \iff \\ 0.17 \leq Z_n \leq 0.5$$

$$P(S_n = 19) \approx P(0.17 \leq Z \leq 0.5)$$

$$= P(Z \leq 0.5) - P(Z \leq 0.17)$$

$$= 0.6915 - 0.5675$$

$$= 0.124$$

- Exact answer:

$$\binom{36}{19} \left(\frac{1}{2}\right)^{36} = 0.1251$$

Poisson vs. normal approximations of the binomial

- Poisson arrivals during unit interval equals: sum of n (independent) Poisson arrivals during n intervals of length $1/n$
 - Let $n \rightarrow \infty$, apply CLT (???)
 - Poisson=normal (????)
- Binomial(n, p)
 - p fixed, $n \rightarrow \infty$: normal
 - np fixed, $n \rightarrow \infty, p \rightarrow 0$: Poisson
- $p = 1/100, n = 100$: Poisson
- $p = 1/10, n = 500$: normal

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013

Transcript – Lecture 21

The following content is provided under a Creative Commons license. Your support will help MIT OpenCourseWare continue to offer high quality educational resources for free. To make a donation or view additional materials from hundreds of MIT courses, visit MIT OpenCourseWare at ocw.mit.edu.

PROFESSOR: It involves real phenomena out there. So we have real stuff that happens. So it might be an arrival process to a bank that we're trying to model.

This is a reality, but this is what we have been doing so far. We have been playing with models of probabilistic phenomena. And somehow we need to tie the two together.

The way these are tied is that we observe the real world and this gives us data. And then based on these data, we try to come up with a model of what exactly is going on. For example, for an arrival process, you might ask the model in question, is my arrival process Poisson or is it something different? If it is Poisson, what is the rate of the arrival process? Once you come up with your model and you come up with the parameters of the model, then you can use it to make predictions about reality or to figure out certain hidden things, certain hidden aspects of reality, that you do not observe directly, but you try to infer what they are. So that's where the usefulness of the model comes in.

Now this field is of course tremendously useful. And it shows up pretty much everywhere. So we talked about the polling examples in the last couple of lectures. This is, of course, a real application.

You sample and on the basis of the sample that you have, you try to make some inferences about, let's say, the preferences in a given population. Let's say in the medical field, you want to try whether a certain drug makes a difference or not. So people would do medical trials, get some results, and then from the data somehow you need to make sense of them and make a decision. Is the new drug useful or is it not? How do we go systematically about the question of this type?

A sexier, more recent topic, there's this famous Netflix competition where Netflix gives you a huge table of movies and people. And people have rated the movies, but not everyone has watched all of the movies in there. You have some of the ratings.

For example, this person gave a 4 to that particular movie. So you get the table that's partially filled. And the Netflix asks you to make recommendations to people.

So this means trying to guess. This person here, how much would they like this particular movie? And you can start thinking, well, maybe this person has given somewhat similar ratings with another person.

And if that other person has also seen that movie, maybe the rating of that other person is relevant. But of course it's a lot more complicated than that. And this has been a serious

competition where people have been using every heavy, wet machinery that there is in statistics, trying to come up with good recommendation systems.

Then the other people, of course, are trying to analyze financial data. Somebody gives you the sequence of the values, let's say of the SMP index. You look at something like this and you can ask questions. How do I model these data using any of the models that we have in our bag of tools? How can I make predictions about what's going to happen afterwards, and so on?

On the engineering side, anywhere where you have noise inference comes in. Signal processing, in some sense, is just an inference problem. You observe signals that are noisy and you try to figure out exactly what's happening out there or what kind of signal has been sent.

Maybe the beginning of the field could be traced a few hundred years ago where people would observe, make astronomical observations of the position of the planets in the sky. They would have some beliefs that perhaps the orbits of planets is an ellipse. Or if it's a comet, maybe it's a parabola, hyperbola, don't know what it is. But they would have a model of that.

But, of course, astronomical measurements would not be perfectly exact. And they would try to find the curve that fits these data. How do you go about choosing this particular curve on the base of noisy data and try to do it in a somewhat principled way?

OK, so questions of this type-- clearly the applications are all over the place. But how is this related conceptually with what we have been doing so far? What's the relation between the field of inference and the field of probability as we have been practicing until now?

Well, mathematically speaking, what's going to happen in the next few lectures could be just exercises or homework problems in the class in based on what we have done so far. That means you're not going to get any new facts about probability theory. Everything we're going to do will be simple applications of things that you already do know.

So in some sense, statistics and inference is just an applied exercise in probability. But actually, things are not that simple in the following sense. If you get a probability problem, there's a correct answer.

There's a correct solution. And that correct solution is unique. There's no ambiguity.

The theory of probability has clearly defined rules. These are the axioms. You're given some information about probability distributions.

You're asked to calculate certain other things. There's no ambiguity. Answers are always unique.

In statistical questions, it's no longer the case that the question has a unique answer. If I give you data and I ask you what's the best way of estimating the motion of that planet, reasonable people can come up with different methods. And reasonable people will try to argue that's my method has these desirable properties but somebody else may say, here's another method that has certain desirable properties. And it's not clear what the best method is.

So it's good to have some understanding of what the issues are and to know at least what is the general class of methods that one tries to consider, how does one go about such problems. So we're going to see lots and lots of different inference methods. We're not going to tell you that one is better than the other. But it's important to understand what are the concepts between those different methods.

And finally, statistics can be misused really badly. That is, one can come up with methods that you think are sound, but in fact they're not quite that. I will bring some examples next time and talk a little more about this.

So, they want to say, you have some data, you want to make some inference from them, what many people will do is to go to Wikipedia, find a statistical test that they think it applies to that situation, plug in numbers, and present results. Are the conclusions that they get really justified or are they misusing statistical methods?

Well, too many people actually do misuse statistics and conclusions that people get are often false. So it's important to, besides just being able to copy statistical tests and use them, to understand what are the assumptions between the different methods and what kind of guarantees they have, if any. All right, so we'll try to do a quick tour through the field of inference in this lecture and the next few lectures that we have left this semester and try to highlight at the very high level the main concept skills, and techniques that come in. Let's start with some generalities and some general statements.

One first statement is that statistics or inference problems come up in very different guises. And they may look as if they are of very different forms. Although, at some fundamental level, the basic issues turn out to be always pretty much the same.

So let's look at this example. There's an unknown signal that's being sent. It's sent through some medium, and that medium just takes the signal and amplifies it by a certain number.

So you can think of somebody shouting. There's the air out there. What you shouted will be attenuated through the air until it gets to a receiver. And that receiver then observes this, but together with some random noise.

Here I meant S. S is the signal that's being sent. And what you observe is an X.

You observe X, so what kind of inference problems could we have here? In some cases, you want to build a model of the physical phenomenon that you're dealing with. So for example, you don't know the attenuation of your signal and you try to find out what this number is based on the observations that you have.

So the way this is done in engineering systems is that you design a certain signal, you know what it is, you shout a particular word, and then the receiver listens. And based on the intensity of the signal that they get, they try to make a guess about A. So you don't know A, but you know S. And by observing X, you get some information about what A is.

So in this case, you're trying to build a model of the medium through which your signal is propagating. So sometimes one would call problems of this kind, let's say, system identification. In a different version of an inference problem that comes with this picture, you've done your modeling.

You know your A. You know the medium through which the signal is going, but it's a communication system. This person is trying to communicate something to that person. So you send the signal S, but that person receives a noisy version of S. So that person tries to reconstruct S based on X.

So in both cases, we have a linear relation between X and the unknown quantity. In one version, A is the unknown and we know S. In the other version, A is known, and so we try to infer S.

Mathematically, you can see that this is essentially the same kind of problem in both cases. Although, the kind of practical problem that you're trying to solve is a little different. So we will not be making any distinctions between problems of the model building type as opposed to models where you try to estimate some unknown signal and so on. Because conceptually, the tools that one uses for both types of problems are essentially the same.

OK, next a very useful classification of inference problems-- the unknown quantity that you're trying to estimate could be either a discrete one that takes a small number of values. So this could be discrete problems, such as the airplane radar problem we encountered back a long time ago in this class. So there's two possibilities-- an airplane is out there or an airplane is not out there.

And you're trying to make a decision between these two options. Or you can have other problems would you have, let's say, four possible options. You don't know which one is true, but you get data and you try to figure out which one is true.

In problems of these kind, usually you want to make a decision based on your data. And you're interested in the probability of making a correct decision. You would like that probability to be as high as possible.

Estimation problems are a little different. Here you have some continuous quantity that's not known. And you try to make a good guess of that quantity. And you would like your guess to be as close as possible to the true quantity.

So the polling problem was of this type. There was an unknown fraction f of the population that had some property. And you try to estimate f as accurately as you can.

So the distinction here is that usually here the unknown quantity takes on discrete set of values. Here the unknown quantity takes a continuous set of values. Here we're interested in the probability of error.

Here we're interested in the size of the error. Broadly speaking, most inference problems fall either in this category or in that category. Although, if you want to complicate life, you can also think or construct problems where both of these aspects are simultaneously present.

OK, finally since we're in classification mode, there is a very big, important dichotomy into how one goes about inference problems. And here there's two fundamentally different philosophical points of view, which is how do we model the quantity that is unknown?

In one approach, you say there's a certain quantity that has a definite value. It just happens that they don't know it. But it's a number. There's nothing random about it. So think of trying to estimate some physical quantity.

You're making measurements, you try to estimate the mass of an electron, which is a sort of universal physical constant. There's nothing random about it. It's a fixed number. You get data, because you have some measuring apparatus.

And that measuring apparatus, depending on what that results that you get are affected by the true mass of the electron, but there's also some noise. You take the data out of your measuring apparatus and you try to come up with some estimate of that quantity theta. So this is definitely a legitimate picture, but the important thing in this picture is that this theta is written as lowercase. And that's to make the point that it's a real number, not a random variable.

There's a different philosophical approach which says, well, anything that I don't know I should model it as a random variable. Yes, I know. The mass of the electron is not really random. It's a constant.

But I don't know what it is. I have some vague sense, perhaps, what it is perhaps because of the experiments that some other people carried out. So perhaps I have a prior distribution on the possible values of Theta.

And that prior distribution doesn't mean that the nature is random, but it's more of a subjective description of my subjective beliefs of where do I think this constant number happens to be. So even though it's not truly random, I model my initial beliefs before the experiment starts. In terms of a prior distribution, I view it as a random variable. Then I observe another related random variable through some measuring apparatus. And then I use this again to create an estimate.

So these two pictures philosophically are very different from each other. Here we treat the unknown quantities as unknown numbers. Here we treat them as random variables.

When we treat them as a random variables, then we know pretty much already what we should be doing. We should just use the Bayes rule. Based on X, find the conditional distribution of Theta. And that's what we will be doing mostly over this lecture and the next lecture.

Now in both cases, what you end up getting at the end is an estimate. But actually, that estimate is what kind of object is it? It's a random variable in both cases. Why?

Even in this case where theta was a constant, my data are random. I do my data processing. So I calculate a function of the data, the data are random variables.

So out here we output something which is a function of a random variable. So this quantity here will be also random. It's affected by the noise and the experiment that I have been doing.

That's why these estimators will be denoted by uppercase Thetas. And we will be using hats. Hat, usually in estimation, means an estimate of something.

All right, so this is the big picture. We're going to start with the Bayesian version. And then the last few lectures we're going to talk about the non-Bayesian version or the classical one.

By the way, I should say that statisticians have been debating fiercely for 100 years whether the right way to approach statistics is to go the classical way or the Bayesian way. And there have been tides going back and forth between the two sides. These days, Bayesian methods tend to become a little more popular for various reasons. We're going to come back to this later.

All right, so in Bayesian estimation, what we got in our hands is Bayes rule. And if you have Bayes rule, there's not a lot that's left to do. We have different forms of the Bayes rule, depending on whether we're dealing with discrete data, And discrete quantities to estimate, or continuous data, and so on.

In the hypothesis testing problem, the unknown quantity Theta is discrete. So in both cases here, we have a P of Theta. We obtain data, the X's. And on the basis of the X that we observe, we can calculate the posterior distribution of Theta, given the data.

So to use Bayesian inference, what do we start with? We start with some priors. These are our initial beliefs about what Theta that might be. That's before we do the experiment.

We have a model of the experimental apparatus. And the model of the experimental apparatus tells us if this Theta is true, I'm going to see X's of that kind. If that other Theta is true, I'm going to see X's that they are somewhere else. That models my apparatus.

And based on that knowledge, once I observe I have these two functions in my hands, we have already seen that if you know those two functions, you can also calculate the denominator here. So all of these functions are available, so you can compute, you can find a formula for this function as well. And as soon as you observe the data, that X's, you plug in here the numerical value of those X's. And you get a function of Theta. And this is the posterior distribution of Theta, given the data that you have seen.

So you've already done a fair number of exercises of these kind. So we not say more about this. And there's a similar formula as you know for the case where we have continuous data. If the X's are continuous random variable, then the formula is the same, except that X's are described by densities instead of being described by a probability mass functions.

OK, now if Theta is continuous, then we're dealing with estimation problems. But the story is once more the same. You're going to use the Bayes rule to come up with the posterior density of Theta, given the data that you have observed.

Now just for the sake of the example, let's come back to this picture here. Suppose that something is flying in the air, and maybe this is just an object in the air close to the Earth. So because of gravity, the trajectory that it's going to follow it's going to be a parabola.

So this is the general equation of a parabola. Z_t is the position of my objects at time t . But I don't know exactly which parabola it is. So the parameters of the parabola are unknown quantities.

What I can do is to go and measure the position of my objects at different times. But unfortunately, my measurements are noisy. What I want to do is to model the motion of my object. So I guess in the picture, the axis would be t going this way and Z going this way.

And on the basis of the data that they get, these are my X 's. I want to figure out the Θ 's. That is, I want to figure out the exact equation of this parabola.

Now if somebody gives you probability distributions for Θ , these would be your priors. So this is given. We need the conditional distribution of the X 's given the Θ 's.

Well, we have the conditional distribution of Z , given the Θ 's from this equation. And then by playing with this equation, you can also find how is X distributed if Θ takes a particular value.

So you do have all of the densities that you might need. And you can apply the Bayes rule. And at the end, your end result would be a formula for the distribution of Θ , given to the X that you have observed-- except for one sort of computation, or to make things more interesting.

Instead of these X 's and Θ 's being single random variables that we have here, typically those X 's and Θ 's will be multi-dimensional random variables or will correspond to multiple ones. So this little Θ here actually stands for a triplet of Θ_0 , Θ_1 , and Θ_2 . And that X here stands here for the entire sequence of X 's that we have observed.

So in reality, the object that you're going to get at to the end after inference is done is a function that you plug in the values of the data and you get the function of the Θ 's that tells you the relative likelihoods of different Θ triplets.

So what I'm saying is that this is no harder than the problems that you have dealt with so far, except perhaps for the complication that's usually in interesting inference problems. Your Θ 's and X 's are often the vectors of random variables instead of individual random variables.

Now if you are to do estimation in a case where you have discrete data, again the situation is no different. We still have a Bayes rule of the same kind, except that densities gets replaced by PMF's. If X is discrete, you put a P here instead of putting an f .

So an example of an estimation problem with discrete data is similar to the polling problem. You have a coin. It has an unknown parameter Theta. This is the probability of obtaining heads. You flip the coin many times. What can you tell me about the true value of Theta?

A classical statistician, at this point, would say, OK, I'm going to use an estimator, the most reasonable one, which is this. How many heads did they obtain in n trials? Divide by the total number of trials. This is my estimate of the bias of my coin.

And then the classical statistician would continue from here and try to prove some properties and argue that this estimate is a good one. For example, we have the weak law of large numbers that tells us that this particular estimate converges in probability to the true parameter. This is a kind of guarantee that's useful to have. And the classical statistician would pretty much close the subject in this way.

What would the Bayesian person do differently? The Bayesian person would start by assuming a prior distribution of Theta. Instead of treating Theta as an unknown constant, they would say that Theta would speak randomly or pretend that it would speak randomly and assume a distribution on Theta.

So for example, if you don't know they need anything more, you might assume that any value for the bias of the coin is as likely as any other value of the bias of the coin. And this way so the probability distribution that's uniform. Or if you have a little more faith in the manufacturing processes that's created that coin, you might choose your prior to be a distribution that's centered around 1/2 and sits fairly narrowly centered around 1/2.

That would be a prior distribution in which you say, well, I believe that the manufacturer tried to make my coin to be fair. But they often makes some mistakes, so it's going to be, I believe, it's approximately 1/2 but not quite. So depending on your beliefs, you would choose an appropriate prior for the distribution of Theta. And then you would use the Bayes rule to find the probabilities of different values of Theta, based on the data that you have observed.

So no matter which version of the Bayes rule that you use, the end product of the Bayes rule is going to be either a plot of this kind or a plot of that kind. So what am I plotting here? This axis is the Theta axis. These are the possible values of the unknown quantity that we're trying to estimate.

In the continuous case, theta is a continuous random variable. I obtain my data. And I plot for the posterior probability distribution after observing my data. And I'm plotting here the probability density for Theta. So this is a plot of that density.

In the discrete case, theta can take finitely many values or a discrete set of values. And for each one of those values, I'm telling you how likely is that the value to be the correct one, given the data that I have observed. And in general, what you would go back to your boss and report after you've done all your inference work would be either a plot of this kinds or of that kind.

So you go to your boss who asks you, what is the value of Theta? And you say, well, I only have limited data. That I don't know what it is. It could be this, with so much probability. There's probability.

OK, let's throw in some numbers here. There's probability 0.3 that Theta is this value. There's probability 0.2 that Theta is this value, 0.1 that it's this one, 0.1 that it's this one, 0.2 that it's that one, and so on.

OK, now bosses often want simple answers. They say, OK, you're talking too much. What do you think Theta is? And now you're forced to make a decision. If that was the situation and you have to make a decision, how would you make it? Well, I'm going to make a decision that's most likely to be correct. If I make this decision, what's going to happen?

Theta is this value with probability 0.2, which means there's probably 0.8 that they make an error if I make that guess. If I make that decision, this decision has probably 0.3 of being the correct one. So I have probably of error 0.7.

So if you want to just maximize the probability of giving the correct decision, or if you want to minimize the probability of making an incorrect decision, what you're going to choose to report is that value of Theta for which the probability is highest. So in this case, I would choose to report this particular value, the most likely value of Theta, given what I have observed. And that value is called them maximum a posteriori probability estimate. It's going to be this one in our case.

So picking the point in the posterior PMF that has the highest probability. That's the reasonable thing to do. This is the optimal thing to do if you want to minimize the probability of an incorrect inference. And that's what people do usually if they need to report a single answer, if they need to report a single decision.

How about in the estimation context? If that's what you know about Theta, Theta could be around here, but there's also some sharp probability that it is around here. What's the single answer that you would give to your boss?

One option is to use the same philosophy and say, OK, I'm going to find the Theta at which this posterior density is highest. So I would pick this point here and report this particular Theta. So this would be my Theta, again, Theta MAP, the Theta that has the highest a posteriori probability, just because it corresponds to the peak of the density.

But in this context, the maximum a posteriori probability theta was the one that was most likely to be true. In the continuous case, you cannot really say that this is the most likely value of Theta. In a continuous setting, any value of Theta has zero probability, so when we talk about densities. So it's not the most likely. It's the one for which the density, so the probabilities of that neighborhoods, are highest. So the rationale for picking this particular estimate in the continuous case is much less compelling than the rationale that we had in here.

So in this case, reasonable people might choose different quantities to report. And the very popular one would be to report instead the conditional expectation. So I don't know quite what Theta is.

Given the data that I have, Theta has this distribution. Let me just report the average over that distribution. Let me report to the center of gravity of this figure.

And in this figure, the center of gravity would probably be somewhere around here. And that would be a different estimate that you might choose to report. So center of gravity is something around here. And this is a conditional expectation of Theta, given the data that you have.

So these are two, in some sense, fairly reasonable ways of choosing what to report to your boss. Some people might choose to report this. Some people might choose to report that. And a priori, if there's no compelling reason why one would be preferable than other one, unless you set some rules for the game and you describe a little more precisely what your objectives are.

But no matter which one you report, a single answer, a point estimate, doesn't really tell you the whole story. There's a lot more information conveyed by this posterior distribution plot than any single number that you might report. So in general, you may wish to convince your boss that's it's worth their time to look at the entire plot, because that plot sort of covers all the possibilities. It tells your boss most likely we're in that range, but there's also a distinct change that our Theta happens to lie in that range.

All right, now let us try to perhaps differentiate between these two and see under what circumstances this one might be the better estimate to perform. Better with respect to what? We need some rules. So we're going to throw in some rules.

As a warm up, we're going to deal with the problem of making an estimation if you had no information at all, except for a prior distribution. So this is a warm up for what's coming next, which would be estimation that takes into account some information.

So we have a Theta. And because of your subjective beliefs or models by others, you believe that Theta is uniformly distributed between, let's say, 4 and 10. You want to come up with a point estimate.

Let's try to look for an estimate. Call it c , in this case. I want to pick a number with which to estimate the value of Theta. I will be interested in the size of the error that I make. And I really dislike large errors, so I'm going to focus on the square of the error that they make.

So I pick c . Theta that has a random value that I don't know. But whatever it is, once it becomes known, it results into a squared error between what it is and what I guessed that it was. And I'm interested in making a small air on the average, where the average is taken with respect to all the possible and unknown values of Theta.

So the problem, this is a least squares formulation of the problem, where we try to minimize the least squares errors. How do you find the optimal c ? Well, we take that expression and expand it.

And it is, using linearity of expectations-- square minus 2c expected Theta plus c squared-- that's the quantity that we want to minimize, with respect to c.

To do the minimization, take the derivative with respect to c and set it to 0. So that differentiation gives us from here minus 2 expected value of Theta plus 2c is equal to 0. And the answer that you get by solving this equation is that c is the expected value of Theta.

So when you do this optimization, you find that the optimal estimate, the things you should be reporting, is the expected value of Theta. So in this particular example, you would choose your estimate c to be just the middle of these values, which would be 7.

OK, and in case your boss asks you, how good is your estimate? How big is your error going to be? What you could report is the average size of the estimation error that you are making. We picked our estimates to be the expected value of Theta. So for this particular way that I'm choosing to do my estimation, this is the mean squared error that I get. And this is a familiar quantity. It's just the variance of the distribution.

So the expectation is that best way to estimate a quantity, if you're interested in the mean squared error. And the resulting mean squared error is the variance itself. How will this story change if we now have data as well? Now having data means that we can compute posterior distributions or conditional distributions. So we get transported into a new universe where instead of working with the original distribution of Theta, the prior distribution, now we work with the condition of distribution of Theta, given the data that we have observed.

Now remember our old slogan that conditional models and conditional probabilities are no different than ordinary probabilities, except that we live now in a new universe where the new information has been taken into account. So if you use that philosophy and you're asked to minimize the squared error but now that you live in a new universe where X has been fixed to something, what would the optimal solution be? It would again be the expectation of theta, but which expectation? It's the expectation which applies in the new conditional universe in which we live right now.

So because of what we did before, by the same calculation, we would find that the optimal estimates is the expected value of X of Theta, but the optimal estimate that takes into account the information that we have. So the conclusion, once you get your data, if you want to minimize the mean squared error, you should just report the conditional estimation of this unknown quantity based on the data that you have.

So the picture here is that Theta is unknown. You have your apparatus that creates measurements. So this creates an X. You take an X, and here you have a box that does calculations. It does calculations and it spits out the conditional expectation of Theta, given the particular data that you have observed.

And what we have done in this class so far is, to some extent, developing the computational tools and skills to do with this particular calculation-- how to calculate the posterior density for Theta and how to calculate expectations, conditional expectations. So in principle, we know how to do

this. In principle, we can program a computer to take the data and to spit out condition expectations.

Somebody who doesn't think like us might instead design a calculating machine that does something differently and produces some other estimate. So we went through this argument and we decided to program our computer to calculate conditional expectations. Somebody else came up with some other crazy idea for how to estimate the random variable. They came up with some function g and the programmed it, and they designed a machine that estimates Θ 's by outputting a certain g of X .

That could be an alternative estimator. Which one is better? Well, we convinced ourselves that this is the optimal one in a universe where we have fixed the particular value of the data. So what we have proved so far is a relation of this kind. In this conditional universe, the mean squared error that I get-- I'm the one who's using this estimator-- is less than or equal than the mean squared error that this person will get, the person who uses that estimator.

For any particular value of the data, I'm going to do better than the other person. Now the data themselves are random. If I average over all possible values of the data, I should still be better off. If I'm better off for any possible value X , then I should be better off on the average over all possible values of X .

So let us average both sides of this quantity with respect to the probability distribution of X . If you want to do it formally, you can write this inequality between numbers as an inequality between random variables. And it tells that no matter what that random variable turns out to be, this quantity is better than that quantity. Take expectations of both sides, and you get this inequality between expectations overall.

And this last inequality tells me that the person who's using this estimator who produces estimates according to this machine will have a mean squared estimation error that's less than or equal to the estimation error that's produced by the other person. In a few words, the conditional expectation estimator is the optimal estimator. It's the ultimate estimating machine. That's how you should solve estimation problems and report a single value. If you're forced to report a single value and if you're interested in estimation errors.

OK, while we could have told you that story, of course, a month or two ago, this is really about interpretation -- about realizing that conditional expectations have a very nice property. But other than that, any probabilistic skills that come into this business are just the probabilistic skills of being able to calculate conditional expectations, which you already know how to do.

So conclusion, all of optimal Bayesian estimation just means calculating and reporting conditional expectations. Well, if the world were that simple, then statisticians wouldn't be able to find jobs if life is that simple. So real life is not that simple. There are complications. And that perhaps makes their life a little more interesting.

OK, one complication is that we would deal with the vectors instead of just single random variables. I use the notation here as if X was a single random variable. In real life, you get

several data. Does our story change? Not really, same argument-- given all the data that you have observed, you should still report the conditional expectation of Theta.

But what kind of work does it take in order to report this conditional expectation? One issue is that you need to cook up a plausible prior distribution for Theta. How do you do that? In a given application , this is a bit of a judgment call, what prior would you be working with. And there's a certain skill there of not making silly choices.

A more pragmatic, practical issue is that this is a formula that's extremely nice and compact and simple that you can write with minimal ink. But the behind it there could be hidden a huge amount of calculation. So doing any sort of calculations that involve multiple random variables really involves calculating multi-dimensional integrals.

And the multi-dimensional integrals are hard to compute. So implementing actually this calculating machine here may not be easy, might be complicated computationally. It's also complicated in terms of not being able to derive intuition about it. So perhaps you might want to have a simpler version, a simpler alternative to this formula that's easier to work with and easier to calculate.

We will be talking about one such simpler alternative next time. So again, to conclude, at the high level, Bayesian estimation is very, very simple, given that you have mastered everything that has happened in this course so far. There are certain practical issues and it's also good to be familiar with the concepts and the issues that in general, you would prefer to report that complete posterior distribution. But if you're forced to report a point estimate, then there's a number of reasonable ways to do it. And perhaps the most reasonable one is to just the report the conditional expectation itself.

MIT OpenCourseWare
<http://ocw.mit.edu>

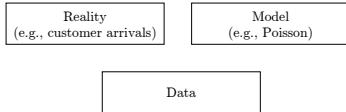
6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

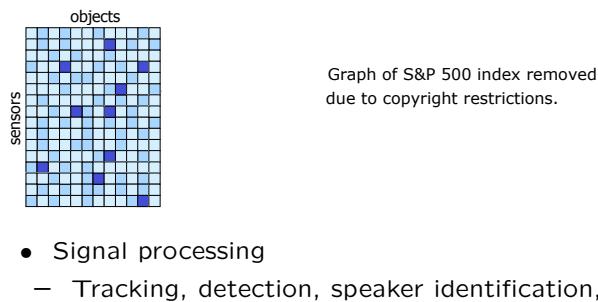
LECTURE 21

- **Readings:** Sections 8.1-8.2

'It is the mark of truly educated people to be deeply moved by **statistics**.'
(Oscar Wilde)



- Design & interpretation of experiments
 - polling, medical/pharmaceutical trials...
- Netflix competition • Finance

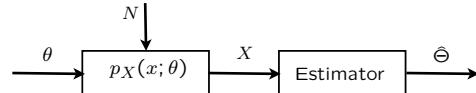


- Signal processing
 - Tracking, detection, speaker identification,...

Types of Inference models/approaches

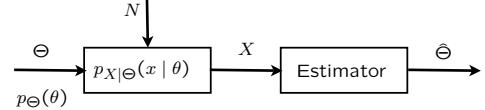
- Model building versus inferring unknown variables. E.g., assume $X = aS + W$
 - Model building:
know "signal" S , observe X , infer a
 - Estimation in the presence of noise:
know a , observe X , estimate S .
- **Hypothesis testing:** unknown takes one of few possible values; aim at small probability of incorrect decision
- **Estimation:** aim at a small estimation error

Classical statistics:



θ : unknown parameter (not a r.v.)
o E.g., θ = mass of electron

- **Bayesian:** Use priors & Bayes rule



Bayesian inference: Use Bayes rule

- **Hypothesis testing**

- discrete data

$$p_{\Theta|X}(\theta | x) = \frac{p_{\Theta}(\theta) p_{X|\Theta}(x | \theta)}{p_X(x)}$$

- continuous data

$$p_{\Theta|X}(\theta | x) = \frac{p_{\Theta}(\theta) f_{X|\Theta}(x | \theta)}{f_X(x)}$$

- **Estimation;** continuous data

$$f_{\Theta|X}(\theta | x) = \frac{f_{\Theta}(\theta) f_{X|\Theta}(x | \theta)}{f_X(x)}$$

$$Z_t = \Theta_0 + t\Theta_1 + t^2\Theta_2$$

$$X_t = Z_t + W_t, \quad t = 1, 2, \dots, n$$

Bayes rule gives:

$$f_{\Theta_0, \Theta_1, \Theta_2 | X_1, \dots, X_n}(\theta_0, \theta_1, \theta_2 | x_1, \dots, x_n)$$

Estimation with discrete data

$$f_{\Theta|X}(\theta | x) = \frac{f_{\Theta}(\theta) p_{X|\Theta}(x | \theta)}{p_X(x)}$$

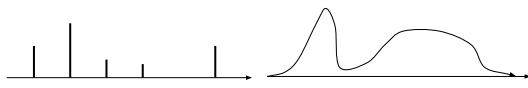
$$p_X(x) = \int f_{\Theta}(\theta) p_{X|\Theta}(x | \theta) d\theta$$

- **Example:**

- Coin with unknown parameter θ
- Observe X heads in n tosses
- What is the Bayesian approach?
- Want to find $f_{\Theta|X}(\theta | x)$
- Assume a prior on Θ (e.g., uniform)

Output of Bayesian Inference

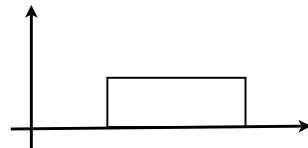
- Posterior distribution:
 - pmf $p_{\Theta|X}(\cdot | x)$ or pdf $f_{\Theta|X}(\cdot | x)$



- If interested in a single answer:
 - Maximum a posteriori probability (MAP):
 - $p_{\Theta|X}(\theta^* | x) = \max_{\theta} p_{\Theta|X}(\theta | x)$ minimizes probability of error; often used in hypothesis testing
 - $f_{\Theta|X}(\theta^* | x) = \max_{\theta} f_{\Theta|X}(\theta | x)$
 - Conditional expectation:
$$E[\Theta | X = y] = \int \theta f_{\Theta|X}(\theta | x) d\theta$$
- Single answers can be misleading!

Least Mean Squares Estimation

- Estimation in the absence of information



- find estimate c , to:

$$\text{minimize } E[(\Theta - c)^2]$$

- Optimal estimate: $c = E[\Theta]$
- Optimal mean squared error:

$$E[(\Theta - E[\Theta])^2] = \text{Var}(\Theta)$$

LMS Estimation of Θ based on X

- Two r.v.'s Θ, X
- we observe that $X = x$
 - new universe: condition on $X = x$
- $E[(\Theta - c)^2 | X = x]$ is minimized by
 $c =$
- $E[(\Theta - E[\Theta | X = x])^2 | X = x]$

$$\leq E[(\Theta - g(x))^2 | X = x]$$
 - $E[(\Theta - E[\Theta | X])^2 | X] \leq E[(\Theta - g(X))^2 | X]$
 - $E[(\Theta - E[\Theta | X])^2] \leq E[(\Theta - g(X))^2]$

$E[\Theta | X]$ minimizes $E[(\Theta - g(X))^2]$
over all estimators $g(\cdot)$

LMS Estimation w. several measurements

- Unknown r.v. Θ
- Observe values of r.v.'s X_1, \dots, X_n
- Best estimator: $E[\Theta | X_1, \dots, X_n]$
- Can be hard to compute/implement
 - involves multi-dimensional integrals, etc.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013 Transcript – Tutorial: An Inference Example

Hi. In this session, we're going to cover a nice review problem that will look at how to infer one random variable based on another. And in this problem, we're given two random variables-- X and Y -- and we're also given their joint pdf, which we're told is a constant $2/3$ within the region bounded by these orange lines. And outside of the region, the joint pdf is 0.

So the first thing we're going to look at, or the first thing that we're asked to do, is find the LMS estimator of Y based on X . Now, remember that LMS estimator is really just a conditional expectation. So the LMS estimator of Y based on X is the conditional expectation of Y , given X .

Now, when we have a plot of the joint pdf and we're dealing with these two random variables, and especially when the joint pdf is constant like this, it's often easy to calculate this conditional expectation visually. So what we really need to do is just say, given any particular value of X , what is the conditional expectation of Y ? So what we can do is we can just pick some values of X and see visually what that initial expectation is.

So for example, if X is $1/2$, given that X is $1/2$, and since this whole joint pdf is uniform, then the conditional slice of Y will be from here to here. And that slice, the conditional distribution of Y , given that X is $1/2$, will also be uniform. So it'll be uniform from here to here. And if it's uniform, we know that the conditional expectation will just be the midpoint here.

And so that would be what the conditional expectation of Y would be, given that X is $1/2$. And we could do the same thing for X equals 1. And we'll see that again, because everything is uniform, this slice is also going to be uniform. And so the conditional expectation will again be the midpoint, which is there.

And then if we just look at it within this region, it's always going to be the midpoint. And so we get that the initial expectation of Y , given X , will just look like that line, which you can think of it as just bisecting this angle formed by these two parts of the region.

But things are a little bit different, though, when we move to the region where X is between 1 and 2. Between 1 and 2, say at 1 and $1/2$, this line doesn't continue. Because now, the slice of Y goes from here to here, and again, it's still uniform. So the midpoint would be there. And similarly for X equals 2, it would be here.

And so for X between 1 and 2, the conditional expectation actually looks like this. So you see that there's actually two linear parts of it, but there's a kink at X equals 1. And so by looking at this visually and taking advantage of the fact that everything is uniform, we can pretty easily figure out what this conditional expectation is.

So now, let's actually just write it out algebraically. So for X between 0 and 1, we said that it's this line, which if we look at it, that's just $1/2$ of X . Now, this is for X between 0 and 1. And if X

is between 1 and 2, it's going to be this line, which is a slope of 1. And if we extend this down, it hits the y-axis at negative 1/2.

So it's $X - 1/2$, if X is between 1 and 2. And otherwise, it's undefined. So we'll focus on these two cases here.

Now, the second part of the question, now that we know what the LMS estimator is, we're asked to find what is the traditional mean squared error of this estimator? So we want to know how good is it. And one way of capturing that is to look at the mean squared error. And so recall that the conditional mean squared error is given by this expression.

So what we're saying is this is what we estimate Y to be. This is what y really is, so this difference is how wrong, or the error in our estimate. We square it, because otherwise, positive and negative errors might cancel each other out, so we square it. And then this just looking at each individual value of x for now. So this is why it's the conditional mean squared error.

So how do we calculate this? Well, remember that this g of X , we said the LMS estimator is just a conditional expectation. So it's just expectation of Y , given X .

Well, then if you look at this, what this reminds you of, it reminds you of the definition of what a conditional variance is. A variance is just, you take the random variable, subtract its mean, square it, and take the expectation of that. This is no different, except that everything is now the conditional world of X .

So this is actually the conditional variance of Y , given X is little x . What is the conditional variance of Y , given that X is little x ? Now, we can again go back to this plot to try to help us out.

We can split this up into regions again. So just take some little x as an example and see what the variance is. So if little x is 1/2, then we know that the conditional distribution of Y would be uniform, we said, from 0 up to here. Well, that point is this from 0 to 1/2.

And remember, the variance of a uniform distribution is just the width of the uniform distribution squared, divided by 12. And so in this case, the width would be 1/2 squared over 12. And in general, for the region of X between 0 and 1, the width of the conditional distribution of Y will always be X , because the width will go from 0 to wherever X is. So because of that, the conditional variance will just be X squared, the width squared, over 12, when X is between 0 and 1.

Now, let's think about the other case, where X is between 1 and 2. Well, if X is between 1 and 2, we're over here. And now, if we take the conditional distribution of Y , it's again uniform. But the width now, instead of varying with Y , it's always going to be the same width. Each of these slices have the same width, and the width goes from here-- this is $X - 1$, and that's X .

So if the width is always going to be a constant of 1. And so this variance is going to be 1/12. And from that, we get our answer for the conditional mean squared error. Now, part c asks us to

find the mean squared error, which is given by this expression. And we'll see that it looks very similar to this, which was the conditional mean squared error.

And now, given what we know from part b, this is easy to calculate. We can just apply total expectation, because this is just equal to the integral of the conditional mean squared error. And then we need to also multiply this by the pf of x, and then integrate over X. And that integral will should be from X equals 0 to 2, because that's the only range that applies for x, given this joint pdf.

Now, in order to do this first, though, we need to figure out what the pdf of X is. In order to do that, we can go back to our original joint pdf of X and Y and marginalize it. So marginalizing, you could think of it as taking this joint pdf and collapsing it onto the x-axis so that you take everything and integrate out Y.

Now to do that, let's do that up here. We can split it up into two sections. So the section of X between 0 and 1, we integrate the joint pdf from Y equals 0 to Y equals X, which is this portion of this line.

So we integrate Y. The joint pdf is $2/3$, and we integrate Y out from Y equals 0 to Y equals X, which is this portion of this line. And then for the portion of X from 1 to 2, we again integrate Y out. Now we integrate Y from X minus 1 up to X.

So this is X between 0 and 1, and this is X between 1 and 2. So we just do some little bit of calculus, and we get that this is going to be $2/3 X$ when X is between 0 and 1. And it's going to be $2/3$ when X is between 1 and 2.

So now that we have what the marginal pdf of X is, we can plug that into this, and plug in what we had for b, and then calculate what this actually is. So remember, we need to take care of these two cases, these two regions-- X between 0 and 1, and X between 1 and 2.

So the conditional mean squared error for X between 0 and 1 is X squared over 2. So between 0 and 1, this first part is X squared over 12. The pdf of X in that same region is $2/3 x$. And we integrate that in the region from x equals 0 to 1.

And then, we also have the second region which is X from 1 to 2. In that region, the traditional mean squared error from part b is $1/12$. The marginal pdf of X is $2/3$, and we do this integral. And if you just carry out some calculus here, you'll get that the final answer is equal to $5/72$.

Now, the last part of this question asks us, is this mean squared error the same thing-- does it equal the expectation of the conditional variance? And it turns out that yes, it does.

And to see that, we can just take this, and use the law of iterated expectations, because iterated expectations tells us this is in fact equal to the expectation of Y minus g of X squared, given X. That's just applying law of iterated expectations. And then, if we look at this, this part that's inside is exactly equal to the conditional variance of Y, given X. And so these two are, in fact, the same.

In part c, we calculated what the marginal pdf of X is, and it'll actually be used later on in this problem. So for future reference, let's just write it down here in this corner. Now, so far in this problem, we've looked at the LMS estimator. And of course, there are other estimators that you can use as well.

And now in part d, let's look at the linear LMS estimator. Now remember, the linear LMS estimator is special, because it forces the estimator to have a linear relationship. So the estimator is going to be a linear function of X .

Now, compare that to what the LMS estimator was in this case. It was two linear pieces, but there was a kink. And so the entire estimator wasn't actually linear in X . Now, the LLMS estimator, or the linear LMS estimator, will give us the linear estimator. It's going to be a linear function of X .

And we know that we have a formula for this. Is the expectation of Y plus the covariance of X and Y over the variance of X times X minus expectation of X . All right, so in order to calculate what this is, we just need to calculate what four things are.

Now, let's start with this last one, the expected value of X . To calculate the expected value of X , we just use a formula. And from before, we know what the pdf of X is. And so we know that this is just going to be X times $f_X(x) dx$.

And in particular, this will give us that from 0 to 1, it's going to be X times the pdf of X is $2/3 X$, so it's $2/3 X^2$. And from 1 to 2, it's going to be X times the pdf of X , which is just $2/3$, so it's $2/3 X$. And when you calculate this out, you'll get that is equal to $11/9$.

Now, let's calculate the variance of X next. In order to calculate that, let's use the formula that variance is equal to the expectation of X^2 minus the expectation of X quantity squared. We had the expectation of X , so let's calculate what the expectation of X^2 is.

Now, it's the same idea. Instead, we have X^2 times $f_X(x) dx$. And we'll get the same sort of formula. We'll split it up again into two different parts from $X = 0$ to $X = 1$. It's going to be X^2 times pdf, so it's $2/3 X^3 dx$.

And then from $X = 1$ to 2 , it's going to be X^2 times $2/3$. So it's $2/3 X^2 dx$. And when we calculate this out, we'll get that it's equal to $31/18$. From that, we know that the variance is going to be equal to expectation of X^2 minus expectation of X quantity squared.

Now, expectation of X^2 is $31/18$. Expectation of X is $11/9$. And when we calculate this, we get that the variance is equal to $37/162$.

So now we have this, and we have that. Let's calculate what expectation of Y is. Expectation of Y , let's calculate it using the law of iterated expectations. The law of iterated expectations tells us that this is equal to the expectation of Y conditioned on X .

Now, we already know what expectation of Y conditioned on X is. That was the LMS estimator that we calculated earlier. It's this. So now we just need to calculate this out, and we can do that.

So we know that in the range from X between 0 and 1, it's equal to $1/2 X$. So in the range from 0 to 1, it's equal to $1/2 X$. But then, we have to use total expectation, so we have to multiply by the pdf of X in that region which is $2/3 X dx$.

And then in the range from X equals 1 to 2, this conditional expectation is $X minus 1/2$. And the pdf of X in that region is $2/3$. Now, when we calculate out this value, we'll get that it's equal to $7/9$.

And now, the last piece is the covariance of X and Y. Remember, the covariance, we can calculate that as the expectation of X times Y minus the expectation of X times the expectation of Y. We already know the expectation of X and the expectation of Y, so we just need to calculate the expectation of X times Y, the product of the two.

And for that, we'll use the definition, and we'll use the joint pdf that we have. So this is going to be a double integral of X times Y times the joint pdf. And the tricky part here is just figuring out what these limits are. So we'll integrate in this order-- X and Y.

Now, let's split this up. So let's focus on splitting X up. So for X between 0 and 1, we just need to figure out what's the rate right range of Y to integrate over such that this is actually non-zero. Because remember, the joint pdf is easy. It's just a constant $2/3$. But it's only a constant $2/3$ within this region.

So the difficult part is just figuring out what the limits are in order to specify that region. So for X between 0 and 1, Y has to be between 0 and X, because this line is Y equals X. So we need to integrate from 0 to X-- X times Y times the joint pdf, which is $2/3$.

And now, let's do the other part, which is X from 1 to 2. Well, if X is from 1 to 2, in order to fall into this region, Y has to be between $X minus 1$ and X. So we integrate Y from $X minus 1$ to X. Against, it's X times Y times the joint pdf, which is $2/3$.

And now, once we have this set up, the rest of it we can just do some calculus. And what we find is that the final answer is equal to $41/36$. Now, what that tells us is that the covariance of X and Y, which is just expectation of X times Y, the product, minus expectation of X times expectation of Y.

We know expectation of X times Y now. It's $41/36$. Expectation of X is $11/9$. Expectation of Y is $7/9$. So when we substitute all of that in, we get that this covariance is $61/324$.

All right, so now we have everything we need. Expectation of Y is here. Covariance is here. Variance of X is here. And expectation of X is here. So let's substitute that in, and we can figure out what the actual LLMS estimator is.

So expectation of Y we know is $7/9$. Expectation of X is $11/9$. And when you divide the covariance, which is $61/324$, by the variance, which is $37/162$, we'll get $61/74$. And so that is the LLMS estimator that we calculated. And notice that it is, in fact, linear in X.

So let's plot that and see what it looks like. So it's going to be a line, and it's going to look like this. So at X equals 2, it's actually a little bit below $1\frac{1}{2}$, which is what the LMS estimator would be. At X equals 1, it's actually a little bit above $1\frac{1}{2}$, which is what the LMS estimator would be. And then it crosses 0 around roughly $1/4$, and it drops actually below 0. So if we connect the dots, it's going to look something like this.

So notice that it's actually not too far away from the LMS estimator here. But it doesn't have the kink because it is a line. And note also that it actually drops below. So when X is very small, you actually estimate negative values of Y, which is actually impossible, given the joint pdf distribution that we're given. And that is sometimes a feature or artifact of the linear LMS estimator, that you'll get values that don't necessarily seem to make sense.

So now that we've calculated the linear LMS estimator in part d, which is this, and the LMS estimator in part a, which is this, we've also compared them visually. The linear LMS estimator is the one in pink, the straight line. And the LMS estimator is the one in black with the kink.

It's an interesting question to now ask, which one of these is better? And in order to judge that, we need to come up with some sort of criterion to compare the two with. And the one that we're going to look at in part e is the mean squared error. Which one gives the lower mean squared error.

And so specifically, we're going to ask ourselves which of these two estimators gives us the smaller mean squared error? Is it the linear LMS estimator given by l of X? Or is it the LMS estimator, given by g of X?

Now, we know that the LMS estimator is the one that actually minimizes this. The LMS estimator is designed to minimize the mean squared error. And so we know that given any estimator of X, this one will have the smallest mean squared error. And so the linear LMS estimator, its mean squared error has to be at least as large as the LMS estimators.

And the last part of the question now asks us to look at a third type of estimator, which is the MEP estimator. Now, we want to ask, why is it that we haven't been using the MEP estimator in this problem? Well, remember what the MEP estimator does.

In this case, what we would do is it would take the conditional distribution ratio of Y given any value of X. And then it would pick the value of Y that gives the highest value in the conditional distribution. And that would be the MEP estimate of Y.

But the problem in this case is that if you take any slice here, so a condition on any value of X, any of these slices, if you just take this out and look at it, it's going to be uniform. This is what the conditional distribution of Y given X is. It's going to be uniform between 0 and X. Now,

what the MEP rule tells us is we're going to pick the value of Y that gives us the highest point in this conditional distribution. You can think of it as a posterior distribution.

Now, what's the problem here? Well, every single point gives us exactly the same value for this conditional distribution. And so there's no unique MEP rule. Every single value of Y has just the same conditional distribution. So there's no sensible way of choosing a value based on the MEP rule in this case.

But compare that with the LMS estimator, which is just get conditional expectation. In that case, we can always find a conditional expectation. In this case, the conditional expectation is the midpoint, which is $X/2$, just as had found in part a.

OK, so in this problem, we reviewed a bunch of different ideas in terms of inference, and we took a joint pdf of X and Y, and we used that to calculate the LMS estimator, the linear LMS estimator. We compared the two, and then we also looked at why in this case, the MEP estimator doesn't really make sense.

All right, so I hope that was helpful, and we'll see you next time.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013 Transcript – Recitation: Inferring a Parameter of Uniform Part 1

Hi. In this problem, Romeo and Juliet are back and they're still looking to meet up for a date. Remember, the last time we met up with them, it was back in the beginning of the course and they were trying to meet up for a date but they weren't always punctual. So we modeled their delay as uniformly distributed between 0 and 1 hour.

So now in this problem, we're actually going to look at variation. And we're going to ask the question, how do we actually know that the distribution is uniformly distributed between 0 and 1 hour? Or it could also be the case that it is uniformly distributed between 0 and half an hour, or zero and two hours. How do we actually know what this parameter of the uniform distribution is?

OK, so let's put ourselves in the shoes of Romeo who's tired of being stood up by Juliet on all these dates. And fortunately, he's learned some probability since the beginning of course, and so have we. And in particular we've learned Bayesian inference. And so in this problem, we're actually going to use basically all the concepts and tools of Bayesian inference that we learned chapter eight and apply them. So it's a nice review problem, and so let's get started.

The set of the problem is similar to the first Romeo and Juliet problem that we dealt with. They are meeting up for a date, and they're not always punctual and they have a delay. But instead of the delay being uniformly distributed between 0 and 1 hour, now we have an extra layer of uncertainty. So if we know sum theta, then we know that the delay, which we'll call x is uniformly distributed between 0 and the theta.

So here's one possible theta, theta 1. But we don't actually know what this theta is. So in the original problem we knew that theta was exactly one hour. But in this problem we don't know what theta is. So theta could also be like this, some other theta 2. And we don't know what this theta is.

And we choose to model it as being uniformly distributed between 0 and 1. So like I said, we have two layers now. We have uncertainty about theta, which is the parameters of the uniform distribution. And then we have uncertainty in regards to the actual delay, x.

OK, so let's actually write out what these distributions are. So theta, the unknown parameter, we're told in the problem that we're going to assume that is uniformly distributed between 0 and 1. And so the PDF is just 1, when theta is between 0 and 1, and 0 otherwise. And we're told that, given what theta is, given what this parameter is, the delay is uniformly distributed between 0 and this theta. So what that means is that we know this conditional PDF, the conditional PDF of x given theta is going to be 1 over theta if x is between 0 and theta, and 0 otherwise.

All right, because we know that given a theta, x is uniformly distributed between 0 and theta. So in order to make this uniform distribution, it's the normalization or the heights, you can think of it, has to be exactly 1 over theta. So just imagine for a concrete case, if theta were 1, 1 hour in the

original problem, then this would just be a PDF of 1 or a standard uniform distribution between 0 and 1. OK, so now this is, we have the necessary fundamentals for this problem.

And what do we do in inference? Well the objective is to try to infer some unknown parameter. And what we have is we have a prior which is our initial belief for what this parameter might be. And then we have some data. So in this case, the data that we collect is the actual observed delayed for Juliet, x .

And this model tells us how this data is essentially generated. And now what we do is, we want to use the data and our prior belief, combined them somehow, and use it to update our belief into what we call our posterior. In order to do that, we use Bayes' rule, which is why this is called Bayesian inference.

So when we use Bayes' rule, remember the Bayes' rule is just, we want to now find the posterior which is the conditional PDF of θ , the unknown parameter, given x . So essentially just flip this condition. And remember Bayes' rule is given as the following. It's just the prior times this conditional PDF of x given θ divided by the PDF of x .

All right, and we know what most of these things are. The prior or just the PDF of θ is 1. The condition PDF of x given θ is 1 over θ . And then of course we have this PDF of x .

But we always have to be careful because these two values are only valid for certain ranges of θ and x . So in order for this to be valid we need θ to be between 0 and 1 because otherwise it would be 0. So we need θ to be between 0 and 1. And we need x to be between 0 and θ . And otherwise this would be 0. So now we're almost done.

One last thing we need to do is just calculate what this denominator is, $f(x)$ of x . Well the denominator, remember, is just a normalization. And it's actually relatively less important because what we'll find out is that this has no dependence on θ . It will only depend on x . So the importance, the dependence on θ , will be captured just by the numerator.

But for completeness let's calculate out what this is. So it's just a normalization. So it's actually just the integral of the numerator. You can think of it as an application of kind of total probability.

So we have this that we integrate over and what do we integrate this over? Well we know that we're integrating over θ . And we know that θ has to be between x -- has to be greater than x and it has to be less than 1. So we integrate from $\theta = x$ to 1.

And this is just the integral from x to 1 of the numerator, right? This is just 1 and this is 1 over θ . So it's the integral of 1 over θ , $d\theta$ from x to 1. Which when you do it out, this is the integral, this is $\log \theta$. So it's $\log 1 - \log x$. Log of 1 is 0.

X , remember x is between 0 and θ . θ is less than 1. So x has to be between 0 and 1.

The log of something between 0 and 1 is negative. So this is a negative number. This is 0. And then we have a negative sign.

So really what we can write this as is the absolute value of log of x. This is just so that it would actually be negative log of x. But because log of x is negative we can just-- we know that this is actually going to be a positive number. So this is just to make it look more intuitive.

OK so now to complete this we can just plug that back in and the final answer is-- this is going to be the absolute value log of x or you could also rewrite this as 1 over theta times absolute value log of x. And of course, remember that the actual limits for where this is valid are very important. OK, so what does this actually mean?

Let's try to interpret what this answer is. So what we have is this is the posterior distribution. And now what have we done? Well we started out with the prior, which was that theta is uniform between 0 and between 0 and 1. This is our prior belief.

Now we observed some data. And this allows us to update our belief. And this is the update that we get.

So let's just assume that we observe that Juliet is late by half an hour. Well if she's late by half an hour, what does that tell us about what theta can be? Well what we know from that at least is that theta cannot be anything less than half an hour because if theta were less than half an hour there's no way that her delay-- remember her delay we know has to be distributed between 0 and theta. There's no way that her delay could be half an hour if theta were less than half an hour. So automatically we know that now theta has to be somewhere between x and one which is where this limit comes in.

So we know that theta have to be between x and 1 now instead of just 0 and 1. So by observing an x that cuts down and eliminates part of the range of theta, the range that theta can take on. Now what else do we know? Well this, we can actually plot this. This is a function of theta.

The log x, we can just think of it as some sort of scaling factor. So it's something like 1 over theta scaled. And so that's going to look something like this. And so what we've done is we've transformed the prior, which looks like flat and uniform into something that looks like this, the posterior.

So we've eliminated small values of x because we know that those can't be possible. And now what's left is everything between x and 1. So now why is it also that it becomes not uniform between x and 1? Well it's because, if you think about it, when theta is close to x, so say x is half an hour. If theta is half an hour, that means that there's higher probability that you will actually observe something, a delay of half an hour because there's only a range between 0 and half an hour that x can be drawn from.

Now if theta was actually 1 then x could be drawn anywhere from 0 to 1 which is a wider range. And so it's less likely that you'll get a value of x equal to half an hour. And so because of that values of theta closer to x are more likely. That's why you get this decreasing function.

OK, so now let's continue and now what we have is this is the case for if you observe one data point. So you arrange a date with Juliet, you observe how late she is, and you get one value of x . And now suppose you want to get collect more data so you arrange say 10 dates with Juliet. And for each one you observe how late she was. So now we can collect multiple samples, say n samples of delays.

So x_1 is her delay on the first date. X_n is her delay on the n th date. And x we can now just call a variable that's a collection of all of these.

And now the question is, how do you incorporate in all this information into updating your belief about θ ? And it's actually pretty analogous to what we've done here. The important assumption that we make in this problem is that conditional on θ , all of these delays are in fact conditionally independent. And that's going to help us solve this problem.

So the set up is essentially the same. What we still need is a-- we still need the prior. And the prior hasn't changed. The prior is still uniform between 0 and 1.

The way the actual delays are generated is we still assume to be the same given conditional on θ , each one of these is conditionally independent, and each one is uniformly distributed between 0 and θ . And so what we get is that this is going to be equal to-- you can also imagine this as a big joint PDF, joint conditional PDF of all the x 's. And because we said that they are conditionally independent given θ , then we can actually split this joint PDF into the product of a lot of individual conditional PDFs. So this we can actually rewrite as PDF of x_1 given θ times all the way through the condition PDF of x_n given θ .

And because we assume that each one of these is-- for each one of these it's uniformly distributed between 0 and θ , they're all the same. So in fact what we get is 1 over θ for each one of these. And there's n of them. So it's 1 over θ to the n .

But what values of x is this valid for? What values of x and θ ? Well what we need is that for each one of these, we need that θ has to be at least equal to whatever x you get. Whatever x you observe, θ has to at least that. So we know that θ has to at least equal to x_1 and all the way through x_n . And so θ has to be at least greater than or equal to all these x 's and otherwise this would be 0.

So let's define something that's going to help us. Let's define \bar{x} to be the maximum of all the observed x 's. And so what we can do is rewrite this condition as θ has to be at least equal to the maximum, equal to \bar{x} . All right, and now we can again apply Bayes' rule. Bayes' rule will tell us what this posterior distribution is.

So again the numerator will be the prior times this conditional PDF over PDF of x . OK, so the numerator again, the prior is just one. This distribution we calculated over here. It's 1 over θ to the n . And then we have this denominator. And again, we need to be careful to write down when this is actually valid. So it's actually valid when \bar{x} is greater than θ -- I'm sorry, \bar{x} is less than or equal to θ , and otherwise it's zero.

So this is actually more or less complete. Again we need to calculate out what exactly this denominator is but just like before it's actually just a scaling factor which is independent of what theta is. So if we wanted to, we could actually calculate this out. It would be just like before. It would be the integral of the numerator, which is 1 over theta to the n d theta. And we integrate theta from before, it was from x to 1.

But now we need to integrate from \bar{x} to 1. And if we wanted to, we can actually do others. It's fairly simple calculus to calculate what this normalization factor would be. But the main point is that the shape of it will be dictated by this 1 over theta to the n term.

And so now we know that with n pieces of data, it's actually going to be 1-- the shape will be 1 over theta to the n, where theta has to be at least greater than or equal to \bar{x} . Before it was actually just 1 over theta and has to be between x and 1. So you can kind of see how the problem generalizes when you collect more data.

So now imagine that this is the new-- when you collect n pieces of data, the maximum of all the x's is here. Well, it turns out that it's the posterior now is going to look something like this. So it becomes steeper because it's 1 over theta to the n as opposed to 1 over theta. And it's limited to be between \bar{x} and 1. And so with more data you're more sure of the range that theta can take on because each data points eliminates parts of theta, the range of theta that theta can't be.

And so you're left with just \bar{x} to 1. And you're also more certain. So you have this kind of distribution.

OK, so this is kind of the posterior distribution which tells you the entire distribution of what the unknown parameter-- the entire distribution of the unknown parameter given all the data that you have plus the prior distribution that you have. But if someone were to come to ask you, your manager asks you, well what is your best guess of what theta is? It's less informative or less clear when you tell them, here's the distribution. Because you still have a big range of what theta could be, it could be anything between x and 1 or \bar{x} and 1. So if you wanted to actually come up with a point estimate which is just one single value, there's different ways you can do it.

The first way that we'll talk about is the map rule. What the map rule does is it takes the posterior distribution and just finds the value of the parameter that gives the maximum posterior distribution, the maximum point in the posterior distribution. So if you look at this posture distribution, the map will just take the highest value.

And in this case, because the posterior looks like this, the highest value is in fact x. And so $\hat{\theta}_{\text{map}}$ is actually just x. And if you think about it, this kind of an optimistic estimate because you always assume that it's whatever, if Juliet were 30 minutes late then you assume that her delay is uniformly distributed between 0 and 30 minutes. Well in fact, even though she arrived 30 minutes late, that could have been because she's actually distributed between 0 and 1 hour and you just happened to get 30 minutes. But what you do is you always take kind of the optimistic, and just give her the benefit of the doubt, and say that was actually kind of the worst case scenario given her distribution.

So another way to take this entire posterior distribution and come up with just a single number, a point estimate, is to take the conditional expectation. So you have an entire distribution. So there's two obvious ways of getting a number out of this. One is to take the maximum and the other is to take the expectation.

So take everything in the distribution, combine it and come up with a estimate. So if you think about it, it will probably be something like here, would be the conditional distribution. So this is called the LMS estimator. And the way to calculate it is just like we said, you take the conditional expectation.

So how do we take the conditional expectation? Remember it is just the value and you weight it by the correct distribution, in this case it's the conditional PDF of theta given x which is the posterior distribution. And what do we integrate theta from? Well we integrate it from x to 1. Now if we plug this in, we integrate from x to 1, theta times the posterior.

The posterior we calculated earlier, it was 1 over theta times the absolute value of log x. So the thetas just cancel out, and you just have 1 over absolute value of log x. Well that doesn't depend on theta. So what you get is just 1 minus x over absolute value of log x.

All right, so we can actually plot this, so we have two estimates now. One is that the estimate is just theta-- the estimate is just x. The other one is that it's 1 minus x over absolute value of log x. So we can plot this and compare the two.

So here's x, and here is theta hat, theta hat of x for the two different estimates. So here's you the estimate from the map rule which is whatever x is, we estimate that theta is equal to x. So it just looks like this. Now if we plot this, turns out that it looks something like this.

And so whatever x is, this will tell you what the estimate, the LMS estimate of theta would be. And it turns out that it's always higher than the map estimate. So it's less optimistic. And it kind of factors in the entire distribution.

So because there are several parts to this problem, we're going to take a pause for a quick break and we'll come back and finish the problem in a little bit.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013 Transcript – Recitation: Inferring a Parameter of Uniform Part 2

Welcome back. So now we're going to finish the rest of this problem. For part e, we've calculated what the map and LMS estimators are. And now we're going to calculate what the conditional mean squared error is. So it's a way to measure how good these estimators are.

So let's start out generically. For any estimator $\hat{\theta}$, the conditional MSE is-- conditional mean squared error-- is equal to this. It's the estimator minus the actual value squared conditioned on X being equal to some little x . So the mean squared error. So you take the error, which is the difference between your estimator and the true value, square it, and then take the mean. And it's conditioned on the actual value of what x is. Or, conditioned on the data that you get is.

So to calculate this, we use our standard definition of what conditional expectation would be. So it's $\hat{\theta}$ minus θ squared. And we weight that by the appropriate conditional PDF, which in this case would be the posterior. And we integrate this from x -- from $\theta = x$ to $\theta = 1$.

Now, we can go through some algebra and this will tell us that this is $\hat{\theta}^2 - 2\hat{\theta}\theta + \theta^2$. And this posterior we know from before is $1/\theta$ times absolute value of $\log x$ $d\theta$.

And when we do out this integral, it's going to be-- we can split up into three different terms. So there's $\hat{\theta}^2$ times this and you integrate it. But in fact, this is just a conditional density. When you integrate it from x to 1, this will just integrate up to 1 because it is a valid density. So the first term is just $\hat{\theta}^2$.

Now, the second term is you can pull out of 2 $\hat{\theta}$ and integrate θ times $1/\theta$ times absolute value of $\log x$ from x to 1.

And then the last one is integral of θ^2 $1/\theta$ times absolute value of $\log x$ $d\theta$ from x to 1. OK, so we can do some more-- with some more calculus, we get a final answer is this. So this will integrate to $1 - x/\log x$. And this will integrate to $1 - x^2/2\log x$.

So this tells us for any generic estimate $\hat{\theta}$, this would be what the conditional mean squared error would be. Now, let's calculate what it actually is for the specific estimates that we actually came up with.

So for the MAP rule, the estimate of $\hat{\theta}$ is just equal to x . So when we plug that into this, we get that the conditional MSE is just equal to $x^2 - 2x + 1 - x/\log x + 1 - x^2/2\log x$.

And for the LMS estimate, remember this was equal to-- theta hat was $1 - \frac{x}{|\log x|}$. And so when you plug this particular theta hat into this formula, what you get is that the conditional mean squared error is equal to $\frac{1 - x^2}{2} + \frac{1}{2} \cdot \frac{1}{x^2}$.

So these two expressions tells us what the mean squared error is for the MAP rule and the LMS rule. And it's kind of hard to actually interpret exactly which one is better based on just these expressions. So it's helpful to plot out what the conditional mean squared error is.

So we're plotting for x . For each possible actual data that we observe-- data point that we observe, what is the mean squared error? So let's do the MAP rule first. The MAP rule would look something like this.

And it turns out that the LMS rule is better, and it will look like this dotted line here on the bottom. And so it turns out that if your metric for how good your estimate is is the conditional mean squared error, then LMS is better than MAP. And this is true because LMS is actually designed to actually minimize what this mean squared error is. And so in this case, the LMS estimator should have a better mean squared error than the map estimator.

OK, now the last part of the problem, we calculate one more type of estimator, which is the linear LMS estimator. So notice that the LMS estimator was this one. It was $1 - \frac{x}{|\log x|}$. And this is not linear in x , which means sometimes it's difficult to calculate. And so what we do is we tried to come up with a linear form of this, something that is like $ax + b$, where a and b are some constant numbers. But that also does well in terms of having a small mean squared error.

And so we know from the class that in order to calculate the linear LMS, the linear LMS we know we just need to calculate a few different parts. So it's equal to the expectation of the parameter plus the covariance of theta and x over the variance of x times x minus expectation of x .

Now, in order to do this, we just need to calculate four things. We need the expectation of theta, the covariance, the variance, and the expectation of x . OK, so let's calculate what these things are.

Expectation of theta. We know that theta is uniformly distributed between 0 and 1. And so the expectation of theta is the easiest one to calculate. It's just $1/2$. What about the expectation of x ?

Well, expectation of x is a little bit more complicated. But remember, like in previous problems, it's helpful when you have a hierarchy of randomness to try to use the law of iterated expectations. So the delay, which is x , is random. But its randomness depends on the actual distribution, which is theta. Which itself is random. And so let's try to condition on theta and see if that helps us. OK, so if we knew what theta was, then what is the expectation of x ?

Well, we know that given theta, x is uniformly distributed between 0 and theta. And so the mean would be just theta over 2. And so this would just be expectation of theta over 2. And we know this is just 1/2 times the expectation of theta, which is 1/2. So this is just 1/4.

Now, let's calculate the variance of x . The variance of x takes some more work because we need to use the law of total variance, which is this. That the variance of theta is equal to the expectation of the conditional variance plus the variance of the conditional expectation.

Let's see if we can figure out what these different parts are. What is the conditional variance of x given theta?

Well, given theta, x we know is uniformly distributed between 0 and theta. And remember for uniform distribution of width c , the variance of that uniform distribution is just c squared over 12. And so in this case, what is the width of this uniform distribution?

Well, it's uniformly distributed between 0 and theta, so the width is theta. So this variance should be theta squared over 12.

OK, what about the expectation of x given theta? Well, we already argued earlier that the expectation of x given theta is just theta over 2. So now let's fill in the rest.

What's the expectation of theta squared over 12? Well, that takes a little bit more work because this is just-- you can think of it as 1/12. You could pull the 1/12 out times the expectation of theta squared. Well, the expectation of theta squared we can calculate from the variance of theta plus the expectation of theta quantity squared. Because that is just the definition of variance. Variance is equal to expectation of theta squared minus expectation of theta quantity squared. So we've just reversed the formula.

Now, the second half is the variance of theta over 2. Well, remember when you pull out a constant from a variance, you have to square it. So this is just equal to 1/4 times the variance of theta.

Well, what is the variance of theta? The variance of theta is the variance of uniform between 0 and 1. So the width is 1. So you get 1 squared over 12. And the variance is 1/12.

What is the mean of theta? It's 1/2 when you square that, you get 1/4. Finally for here, the variance of theta like we said, is 1/12. So you get 1/12.

And now, when you combine all these, you get that the variance ends up being 7/144. Now we have almost everything. The last thing we need to calculate is this covariance term.

What is the covariance of theta and x ? Well, the covariance we know is just the expectation of the product of theta and x minus the product of the expectations. So the expectation of x times the expectation of theta.

All right, so we already know what expectation of theta is. That's 1/2. And expectation of x was 1/4. So the only thing that we don't know is expectation of the product of the two. So once again, let's try to use iterated expectations. So let's calculate this as the expectation of this conditional expectation. So we, again, condition on theta.

And minus the expectation of theta is 1/2. Times 1/4, which is the expectation of x. Now, what is this conditional expectation?

Well, the expectation of theta-- if you know what theta is, then the expectation of theta is just theta. You already know what it is, so you know for sure that the expectation is just equal to theta. And what is the expectation of x given theta?

Well, the expectation of x given theta we already said was theta over 2. So what you get is this entire expression is just going to be equal to theta times theta over 2, or expectation of theta squared over 2 minus 1/8.

Now, what is the expectation of theta squared over 2? Well, we know that-- we already calculated out what expectation of theta squared is. So we know that expectation of theta squared is 1/12 plus 1/4. So what we get is we need a 1/2 times 1/12 plus 1/4, which is 1/3 minus 1/8. So the answer is 1/6 minus 1/8, which is 1/24.

Now, let's actually plug this in and figure out what this value is. So when you get everything-- when you combine everything, you get that the LMS estimator is-- the linear LMS estimator is going to be-- expectation of theta is 1/2. The covariance is 1/24. The variance is 7/144. And when you divide that, it's equal to 6/7 times x minus 1/4 because expectation of x is 1/4.

And you can simplify this a little bit and get that this is equal to 6/7 times x plus 2/7. So now we have three different types of estimators. The map estimator, which is this. Notice that it's kind of complicated. You have x squared terms. You have more x squared terms. And you have absolute value of log of x. And then you have the LMS, which is, again, nonlinear.

And now you have something that looks very simple-- much simpler. It's just 6/7 x plus 2/7. And that is the linear LMS estimator.

And it turns out that you can, again, plot these to see what this one looks like. So here is our original plot of x and theta hat. So the map estimator-- sorry, the map estimator was just theta hat equals x. This was the mean squared error of the map estimator. So the map estimator is just this diagonal straight line.

The LMS estimator looked like this. And it turns out that the linear LMS estimator will look something like this. So it is fairly close to the LMS estimator, but not quite the same.

And note, especially that depending on what x is, if x is fairly close to the 1, you might actually get an estimate of theta that's greater than 1. So for example, if you observe that Julian is actually an hour late, then x is 1 and your estimate of theta from the linear LMS estimator would be 8/7, which is greater than 1.

That doesn't quite make sense because we know that theta is bounded to be only between 0 and 1. So you shouldn't get an estimate of theta that's greater than 1. And that's one of the side effects of having the linear LMS estimator. So that sometimes you will have an estimator that doesn't quite make sense.

But what you get instead when sacrificing that is you get a simple form of the estimator that's linear. And now, let's actually consider what the performance is.

And it turns out that the performance in terms of the conditional mean squared error is actually fairly close to the LMS estimator. So it looks like this. Pretty close, pretty close, until you get close to 1. In which case, it does worse. And it does worse precisely because it will come up with estimates of theta which are greater than 1, which are too large. But otherwise, it does pretty well with a estimator that is much simpler in form than the LMS estimator.

So in this problem, which had several parts, we actually went through, basically, all the different concepts and tools within Chapter Eight for Bayesian inference. We talked about the prior, the posterior, calculating the posterior using the Bayes' rule. We calculated the MAP estimator. We calculated the LMS estimator. From those, we calculated what the mean squared error for each one of those and compared the two.

And then, we looked at the linear LMS estimator as another example and calculated what that estimator is, along with the mean squared error for that and compared all three of these. So I hope that was a good review problem for Chapter Eight, and we'll see you next time.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013 Transcript – Lecture 22

The following content is provided under a Creative Commons license. Your support will help MIT OpenCourseWare continue to offer high quality educational resources for free. To make a donation or view additional materials from hundreds of MIT courses, visit MIT OpenCourseWare at ocw.mit.edu

PROFESSOR: So we're going to finish today our discussion of Bayesian Inference, which we started last time. As you probably saw there's not a huge lot of concepts that we're introducing at this point in terms of specific skills of calculating probabilities. But, rather, it's more of an interpretation and setting up the framework.

So the framework in Bayesian estimation is that there is some parameter which is not known, but we have a prior distribution on it. These are beliefs about what this variable might be, and then we'll obtain some measurements. And the measurements are affected by the value of that parameter that we don't know. And this effect, the fact that X is affected by Θ , is captured by introducing a conditional probability distribution-- the distribution of X depends on Θ . It's a conditional probability distribution.

So we have formulas for these two densities, the prior density and the conditional density. And given that we have these, if we multiply them we can also get the joint density of X and Θ . So we have everything that's there is to know in this second.

And now we observe the random variable X . Given this random variable what can we say about Θ ? Well, what we can do is we can always calculate the conditional distribution of Θ given X . And now that we have the specific value of X we can plot this as a function of Θ .

OK. And this is the complete answer to a Bayesian Inference problem. This posterior distribution captures everything there is to say about Θ , that's what we know about Θ . Given the X that we have observed Θ is still random, it's still unknown. And it might be here, there, or there with several probabilities.

On the other hand, if you want to report a single value for Θ then you do some extra work. You continue from here, and you do some data processing on X . Doing data processing means that you apply a certain function on the data, and this function is something that you design. It's the so-called estimator. And once that function is applied it outputs an estimate of Θ , which we call $\Theta\hat{}$.

So this is sort of the big picture of what's happening. Now one thing to keep in mind is that even though I'm writing single letters here, in general Θ or X could be vector random variables. So think of this-- it could be a collection $\Theta_1, \Theta_2, \Theta_3$. And maybe we obtained several measurements, so this X is really a vector X_1, X_2, \dots, X_n .

All right, so now how do we choose a Theta to report? There are various ways of doing it. One is to look at the posterior distribution and report the value of Theta, at which the density or the PMF is highest. This is called the maximum a posteriori estimate. So we pick a value of theta for which the posteriori is maximum, and we report it. An alternative way is to try to be optimal with respects to a mean squared error. So what is this?

If we have a specific estimator, g , this is the estimate it's going to produce. This is the true value of Theta, so this is our estimation error. We look at the square of the estimation error, and look at the average value. We would like this squared estimation error to be as small as possible. How can we design our estimator g to make that error as small as possible?

It turns out that the answer is to produce, as an estimate, the conditional expectation of Theta given X. So the conditional expectation is the best estimate that you could produce if your objective is to keep the mean squared error as small as possible. So this statement here is a statement of what happens on the average over all Theta's and all X's that may happen in our experiment.

The conditional expectation as an estimator has an even stronger property. Not only it's optimal on the average, but it's also optimal given that you have made a specific observation, no matter what you observe. Let's say you observe the specific value for the random variable X. After that point if you're asked to produce a best estimate Theta hat that minimizes this mean squared error, your best estimate would be the conditional expectation given the specific value that you have observed.

These two statements say almost the same thing, but this one is a bit stronger. This one tells you no matter what specific X happens the conditional expectation is the best estimate. This one tells you on the average, over all X's may happen, the conditional expectation is the best estimator.

Now this is really a consequence of this. If the conditional expectation is best for any specific X, then it's the best one even when X is left random and you are averaging your error over all possible X's.

OK so now that we know what is the optimal way of producing an estimate let's do a simple example to see how things work out. So we have started with an unknown random variable, Theta, which is uniformly distributed between 4 and 10. And then we have an observation model that tells us that given the value of Theta, X is going to be a random variable that ranges between Theta - 1, and Theta + 1. So think of X as a noisy measurement of Theta, plus some noise, which is between -1, and +1.

So really the model that we are using here is that X is equal to Theta plus U -- where U is uniform on -1, and +1. one, and plus one. So we have the true value of Theta, but X could be Theta - 1, or it could be all the way up to Theta + 1. And the X is uniformly distributed on that interval. That's the same as saying that U is uniformly distributed over this interval.

So now we have all the information that we need, we can construct the joint density. And the joint density is, of course, the prior density times the conditional density. We go both of these.

Both of these are constants, so the joint density is also going to be a constant. $1/6$ times $1/2$, this is one over 12. But it is a constant, not everywhere. Only on the range of possible x 's and θ 's. So θ can take any value between four and ten, so these are the values of θ . And for any given value of θ x can take values from θ minus one, up to θ plus one.

So here, if you can imagine, a line that goes with slope one, and then x can take that value of θ plus or minus one. So this object here, this is the set of possible x and θ pairs. So the density is equal to one over 12 over this set, and it's zero everywhere else. So outside here the density is zero, the density only applies at that point.

All right, so now we're asked to estimate θ in terms of x . So we want to build an estimator which is going to be a function from the x 's to the θ 's. That's why I chose the axis this way-- x to be on this axis, θ on that axis-- Because the estimator we're building is a function of x . Based on the observation that we obtained, we want to estimate θ .

So we know that the optimal estimator is the conditional expectation, given the value of x . So what is the conditional expectation? If you fix a particular value of x , let's say in this range. So this is our x , then what do we know about θ ? We know that θ lies in this range. θ can only be sampled between those two values. And what kind of distribution does θ have? What is the conditional distribution of θ given x ?

Well, remember how we built conditional distributions from joint distributions? The conditional distribution is just a section of the joint distribution applied to the place where we're conditioning. So the joint is constant. So the conditional is also going to be a constant density over this interval. So the posterior distribution of θ is uniform over this interval.

So if the posterior of θ is uniform over that interval, the expected value of θ is going to be the meet point of that interval. So the estimate which you report-- if you observe that θ -- is going to be this particular point here, it's the midpoint.

The same argument goes through even if you obtain an x somewhere here. Given this x , θ can take a value between these two values. θ is going to have a uniform distribution over this interval, and the conditional expectation of θ given x is going to be the midpoint of that interval.

So now if we plot our estimator by tracing midpoints in this diagram what you're going to obtain is a curve that starts like this, then it changes slope. So that it keeps track of the midpoint, and then it goes like that again. So this blue curve here is our g of x , which is the conditional expectation of θ given that x is equal to little x .

So it's a curve, in our example it consists of three straight segments. But overall it's non-linear. It's not a single line through this diagram. And that's how things are in general. g of x , our optimal estimate has no reason to be a linear function of x . In general it's going to be some complicated curve.

So how good is our estimate? I mean you reported your x , your estimate of theta based on x , and your boss asks you what kind of error do you expect to get? Having observed the particular value of x , what you can report to your boss is what you think is the mean squared error is going to be. We observe the particular value of x . So we're conditioning, and we're living in this universe.

Given that we have made this observation, this is the true value of theta, this is the estimate that we have produced, this is the expected squared error, given that we have made the particular observation. Now in this conditional universe this is the expected value of theta given x . So this is the expected value of this random variable inside the conditional universe.

So when you take the mean squared of a random variable minus the expected value, this is the same thing as the variance of that random variable. Except that it's the variance inside the conditional universe. Having observed x , theta is still a random variable. It's distributed according to the posterior distribution. Since it's a random variable, it has a variance. And that variance is our mean squared error.

So this is the variance of the posterior distribution of Theta given the observation that we have made. OK, so what is the variance in our example? If X happens to be here, then Theta is uniform over this interval, and this interval has length 2. Theta is uniformly distributed over an interval of length 2. This is the posterior distribution of Theta. What is the variance? Then you remember the formula for the variance of a uniform random variable, it is the length of the interval squared divided by 12, so this is $1/3$.

So the variance of Theta -- the mean squared error-- is going to be $1/3$ whenever this kind of picture applies. This picture applies when X is between 5 and 9. If X is less than 5, then the picture is a little different, and Theta is going to be uniform over a smaller interval. And so the variance of theta is going to be smaller as well.

So let's start plotting our mean squared error. Between 5 and 9 the variance of Theta -- the posterior variance-- is $1/3$. Now when the X falls in here Theta is uniformly distributed over a smaller interval. The size of this interval changes linearly over that range. And so when we take the square size of that interval we get a quadratic function of how much we have moved from that corner.

So at that corner what is the variance of Theta? Well if I observe an X that's equal to 3 then I know with certainty that Theta is equal to 4. Then I'm in very good shape, I know exactly what Theta is going to be. So the variance, in this case, is going to be 0.

If I observe an X that's a little larger than Theta is now random, takes values on a little interval, and the variance of Theta is going to be proportional to the square of the length of that little interval. So we get a curve that starts rising quadratically from here. It goes up forward $1/3$. At the other end of the picture the same is true. If you observe an X which is 11 then Theta can only be equal to 10.

And so the error in Theta is equal to 0, there's 0 error variance. But as we obtain X 's that are slightly less than 11 then the mean squared error again rises quadratically. So we end up with a

plot like this. What this plot tells us is that certain measurements are better than others. If you're lucky, and you see X equal to 3 then you're lucky, because you know Θ exactly what it is.

If you see an X which is equal to 6 then you're sort of unlikely, because it doesn't tell you Θ with great precision. Θ could be anywhere on that interval. And so the variance of Θ -- even after you have observed X -- is a certain number, $1/3$ in our case.

So the moral to keep out of that story is that the error variance-- or the mean squared error-- depends on what particular observation you happen to obtain. Some observations may be very informative, and once you see a specific number than you know exactly what Θ is. Some observations might be less informative. You observe your X , but it could still leave a lot of uncertainty about Θ .

So conditional expectations are really the cornerstone of Bayesian estimation. They're particularly popular, especially in engineering contexts. There used a lot in signal processing, communications, control theory, so on. So that makes it worth playing a little bit with their theoretical properties, and get some appreciation of a few subtleties involved here.

No new math in reality, in what we're going to do here. But it's going to be a good opportunity to practice manipulation of conditional expectations. So let's look at the expected value of the estimation error that we obtained. So Θ hat is our estimator, is the conditional expectation. Θ hat minus Θ is what kind of error do we have? If Θ hat, is bigger than Θ then we have made the positive error.

If not, if it's on the other side, we have made the negative error. Then it turns out that on the average the errors cancel each other out, on the average. So let's do this calculation. Let's calculate the expected value of the error given X . Now by definition the error is expected value of Θ hat minus Θ given X .

We use linearity of expectations to break it up as expected value of Θ hat given X minus expected value of Θ given X . And now what? Our estimate is made on the basis of the data of the X 's.

If I tell you X then you know what Θ hat is. Remember that the conditional expectation is a random variable which is a function of the random variable, on which you're conditioning on. If you know X then you know the conditional expectation given X , you know what Θ hat is going to be.

So Θ hat is a function of X . If it's a function of X then once I tell you X you know what Θ hat is going to be. So this conditional expectation is going to be Θ hat itself. Here this is-- just by definition-- Θ hat, and so we get equality to 0. So what we have proved is that no matter what I have observed, and given that I have observed something on the average my error is going to be 0.

This is a statement involving equality of random variables. Remember that conditional expectations are random variables because they depend on the thing you're conditioning on. 0 is

sort of a trivial random variable. This tells you that this random variable is identically equal to the 0 random variable.

More specifically it tells you that no matter what value for X you observe, the conditional expectation of the error is going to be 0. And this takes us to this statement here, which is inequality between numbers. No matter what specific value for capital X you have observed, your error, on the average, is going to be equal to 0.

So this is a less abstract version of these statements. This is inequality between two numbers. It's true for every value of X , so it's true in terms of these random variables being equal to that random variable. Because remember according to our definition this random variable is the random variable that takes this specific value when capital X happens to be equal to little x .

Now this doesn't mean that your error is 0, it only means that your error is as likely, in some sense, to fall on the positive side, as to fall on the negative side. So sometimes your error will be positive, sometimes negative. And on the average these things cancel out and give you a 0 -- on the average.

So this is a property that's sometimes giving the name we say that $\hat{\Theta}$ is unbiased. So $\hat{\Theta}$, our estimate, does not have a tendency to be on the high side. It does not have a tendency to be on the low side. On the average it's just right.

So let's do a little more playing here. Let's see how our error is related to an arbitrary function of the data. Let's do this in a conditional universe and look at this quantity.

In a conditional universe where X is known then h of X is known. And so you can pull it outside the expectation. In the conditional universe where the value of X is given this quantity becomes just a constant. There's nothing random about it. So you can pull it out, the expectation, and write things this way. And we have just calculated that this quantity is 0. So this number turns out to be 0 as well.

Now having done this, we can take expectations of both sides. And now let's use the law of iterated expectations. Expectation of a conditional expectation gives us the unconditional expectation, and this is also going to be 0. So here we use the law of iterated expectations. OK.

OK, why are we doing this? We're doing this because I would like to calculate the covariance between $\tilde{\Theta}$ and $\hat{\Theta}$. $\hat{\Theta}$ is, ask the question -- is there a systematic relation between the error and the estimate?

So to calculate the covariance we use the property that we can calculate the covariances by calculating the expected value of the product minus the product of the expected values.

And what do we get? This is 0, because of what we just proved. And this is 0, because of what we proved earlier. That the expected value of the error is equal to 0.

So the covariance between the error and any function of X is equal to 0. Let's use that to the case where the function of X we're considering is $\Theta\hat{\theta}$ itself.

$\Theta\hat{\theta}$ is our estimate, it's a function of X . So this 0 result would still apply, and we get that this covariance is equal to 0.

OK, so that's what we proved. Let's see, what are the morals to take out of all this? First is you should be very comfortable with this type of calculation involving conditional expectations. The main two things that we're using are that when you condition on a random variable any function of that random variable becomes a constant, and can be pulled out the conditional expectation.

The other thing that we are using is the law of iterated expectations, so these are the skills involved. Now on the substance, why is this result interesting? This tells us that the error is uncorrelated with the estimate. What's a hypothetical situation where these would not happen? Whenever $\Theta\hat{\theta}$ is positive my error tends to be negative.

Suppose that whenever $\Theta\hat{\theta}$ is big then you say oh my estimate is too big, maybe the true Θ is on the lower side, so I expect my error to be negative. That would be a situation that would violate this condition. This condition tells you that no matter what $\Theta\hat{\theta}$ is, you don't expect your error to be on the positive side or on the negative side. Your error will still be 0 on the average.

So if you obtain a very high estimate this is no reason for you to suspect that the true Θ is lower than your estimate. If you suspected that the true Θ was lower than your estimate you should have changed your $\Theta\hat{\theta}$.

If you make an estimate and after obtaining that estimate you say I think my estimate is too big, and so the error is negative. If you thought that way then that means that your estimate is not the optimal one, that your estimate should have been corrected to be smaller. And that would mean that there's a better estimate than the one you used, but the estimate that we are using here is the optimal one in terms of mean squared error, there's no way of improving it. And this is really captured in that statement. That is knowing $\Theta\hat{\theta}$ doesn't give you a lot of information about the error, and gives you, therefore, no reason to adjust your estimate from what it was.

Finally, a consequence of all this. This is the definition of the error. Send Θ to this side, send $\Theta\tilde{\theta}$ to that side, you get this relation. The true parameter is composed of two quantities. The estimate, and the error that they got with a minus sign. These two quantities are uncorrelated with each other. Their covariance is 0, and therefore, the variance of this is the sum of the variances of these two quantities.

So what's an interpretation of this equality? There is some inherent randomness in the random variable Θ that we're trying to estimate. $\Theta\hat{\theta}$ tries to estimate it, tries to get close to it. And if $\Theta\hat{\theta}$ always stays close to Θ , since Θ is random $\Theta\hat{\theta}$ must also be quite random, so it has uncertainty in it.

And the more uncertain Θ is the more it moves together with $\hat{\Theta}$. So the more uncertainty it removes from Θ . And this is the remaining uncertainty in Θ . The uncertainty that's left after we've done our estimation. So ideally, to have a small error we want this quantity to be small. Which is the same as saying that this quantity should be big.

In the ideal case $\hat{\Theta}$ is the same as Θ . That's the best we could hope for. That corresponds to 0 error, and all the uncertainty in Θ is absorbed by the uncertainty in $\hat{\Theta}$.

Interestingly, this relation here is just another variation of the law of total variance that we have seen at some point in the past. I will skip that derivation, but it's an interesting fact, and it can give you an alternative interpretation of the law of total variance.

OK, so now let's return to our example. In our example we obtained the optimal estimator, and we saw that it was a nonlinear curve, something like this. I'm exaggerating the corner of a little bit to show that it's nonlinear.

This is the optimal estimator. It's a nonlinear function of X -- nonlinear generally means complicated.

Sometimes the conditional expectation is really hard to compute, because whenever you have to compute expectations you need to do some integrals. And if you have many random variables involved it might correspond to a multi-dimensional integration. We don't like this. Can we come up, maybe, with a simpler way of estimating Θ ? Of coming up with a point estimate which still has some nice properties, it has some good motivation, but is simpler. What does simpler mean? Perhaps linear.

Let's put ourselves in a straitjacket and restrict ourselves to estimators that are of these forms. My estimate is constrained to be a linear function of the X 's. So my estimator is going to be a curve, a linear curve. It could be this, it could be that, maybe it would want to be something like this. I want to choose the best possible linear function.

What does that mean? It means that I write my $\hat{\Theta}$ in this form. If I fix a certain a and b I have fixed the functional form of my estimator, and this is the corresponding mean squared error. That's the error between the true parameter and the estimate of that parameter, we take the square of this.

And now the optimal linear estimator is defined as one for which these mean squared error is smallest possible over all choices of a and b . So we want to minimize this expression over all a 's and b 's. How do we do this minimization?

Well this is a square, you can expand it. Write down all the terms in the expansion of the square. So you're going to get the term expected value of Θ squared. You're going to get another term-- a squared expected value of X squared, another term which is b squared, and then you're going to get to various cross terms. What you have here is really a quadratic function of a and b .

So think of this quantity that we're minimizing as some function h of a and b , and it happens to be quadratic.

How do we minimize a quadratic function? We set the derivative of this function with respect to a and b to 0, and then do the algebra. After you do the algebra you find that the best choice for a is this 1, so this is the coefficient next to X . This is the optimal a .

And the optimal b corresponds of the constant terms. So this term and this times that together are the optimal choices of b . So the algebra itself is not very interesting. What is really interesting is the nature of the result that we get here.

If we were to plot the result on this particular example you would get the curve that's something like this. It goes through the middle of this diagram and is a little slanted. In this example, X and Θ are positively correlated. Bigger values of X generally correspond to bigger values of Θ .

So in this example the covariance between X and Θ is positive, and so our estimate can be interpreted in the following way: The expected value of Θ is the estimate that you would come up with if you didn't have any information about Θ . If you don't make any observations this is the best way of estimating Θ .

But I have made an observation, X , and I need to take it into account. I look at this difference, which is the piece of news contained in X ? That's what X should be on the average. If I observe an X which is bigger than what I expected it to be, and since X and Θ are positively correlated, this tells me that Θ should also be bigger than its average value.

Whenever I see an X that's larger than its average value this gives me an indication that Θ should also probably be larger than its average value. And so I'm taking that difference and multiplying it by a positive coefficient. And that's what gives me a curve here that has a positive slope.

So this increment-- the new information contained in X as compared to the average value we expected apriori, that increment allows us to make a correction to our prior estimate of Θ , and the amount of that correction is guided by the covariance of X with Θ . If the covariance of X with Θ were 0, that would mean there's no systematic relation between the two, and in that case obtaining some information from X doesn't give us a guide as to how to change the estimates of Θ .

If that were 0, we would just stay with this particular estimate. We're not able to make a correction. But when there's a non zero covariance between X and Θ that covariance works as a guide for us to obtain a better estimate of Θ .

How about the resulting mean squared error? In this context turns out that there's a very nice formula for the mean squared error obtained from the best linear estimate. What's the story here?

The mean squared error that we have has something to do with the variance of the original random variable. The more uncertain our original random variable is, the more error we're going to make. On the other hand, when the two variables are correlated we explored that correlation to improve our estimate.

This row here is the correlation coefficient between the two random variables. When this correlation coefficient is larger this factor here becomes smaller. And our mean squared error become smaller. So think of the two extreme cases. One extreme case is when rho equal to 1 -- so X and Theta are perfectly correlated.

When they're perfectly correlated once I know X then I also know Theta. And the two random variables are linearly related. In that case, my estimate is right on the target, and the mean squared error is going to be 0.

The other extreme case is if rho is equal to 0. The two random variables are uncorrelated. In that case the measurement does not help me estimate Theta, and the uncertainty that's left-- the mean squared error-- is just the original variance of Theta. So the uncertainty in Theta does not get reduced.

So moral-- the estimation error is a reduced version of the original amount of uncertainty in the random variable Theta, and the larger the correlation between those two random variables, the better we can remove uncertainty from the original random variable.

I didn't derive this formula, but it's just a matter of algebraic manipulations. We have a formula for Theta hat, subtract Theta from that formula. Take square, take expectations, and do a few lines of algebra that you can read in the text, and you end up with this really neat and clean formula.

Now I mentioned in the beginning of the lecture that we can do inference with Theta's and X's not just being single numbers, but they could be vector random variables. So for example we might have multiple data that gives us information about X.

There are no vectors here, so this discussion was for the case where Theta and X were just scalar, one-dimensional quantities. What do we do if we have multiple data? Suppose that Theta is still a scalar, it's one dimensional, but we make several observations. And on the basis of these observations we want to estimate Theta.

The optimal least mean squares estimator would be again the conditional expectation of Theta given X. That's the optimal one. And in this case X is a vector, so the general estimator we would use would be this one.

But if we want to keep things simple and we want our estimator to have a simple functional form we might restrict to estimator that are linear functions of the data. And then the story is exactly the same as we discussed before. I constrained myself to estimating Theta using a linear function of the data, so my signal processing box just applies a linear function.

And I'm looking for the best coefficients, the coefficients that are going to result in the least possible squared error. This is my squared error, this is (my estimate minus the thing I'm trying to estimate) squared, and then taking the average. How do we do this? Same story as before.

The X's and the Theta's get averaged out because we have an expectation. Whatever is left is just a function of the coefficients of the a's and of b's. As before it turns out to be a quadratic function. Then we set the derivatives of this function of a's and b's with respect to the coefficients, we set it to 0.

And this gives us a system of linear equations. It's a system of linear equations that's satisfied by those coefficients. It's a linear system because this is a quadratic function of those coefficients. So to get closed-form formulas in this particular case one would need to introduce vectors, and matrices, and metrics inverses and so on.

The particular formulas are not so much what interests us here, rather, the interesting thing is that this is simply done just using straightforward solvers of linear equations. The only thing you need to do is to write down the correct coefficients of those non-linear equations. And the typical coefficient that you would get would be what? Let say a typical quick equations would be -- let's take a typical term of this quadratic one you expanded.

You're going to get the terms such as a_1x_1 times a_2x_2 . When you take expectations you're left with a_1a_2 times expected value of x_1x_2 . So this would involve terms such as a_1 squared expected value of x_1 squared. You would get terms such as a_1a_2 , expected value of x_1x_2 , and a lot of other terms here should have a too.

So you get something that's quadratic in your coefficients. And the constants that show up in your system of equations are things that have to do with the expected values of squares of your random variables, or products of your random variables. To write down the numerical values for these the only thing you need to know are the means and variances of your random variables. If you know the mean and variance then you know what this thing is. And if you know the covariances as well then you know what this thing is.

So in order to find the optimal linear estimator in the case of multiple data you do not need to know the entire probability distribution of the random variables that are involved. You only need to know your means and covariances. These are the only quantities that affect the construction of your optimal estimator.

We could see this already in this formula. The form of my optimal estimator is completely determined once I know the means, variance, and covariance of the random variables in my model. I do not need to know how the details distribution of the random variables that are involved here.

So as I said in general, you find the form of the optimal estimator by using a linear equation solver. There are special examples in which you can get closed-form solutions. The nicest simplest estimation problem one can think of is the following-- you have some uncertain parameter, and you make multiple measurements of that parameter in the presence of noise.

So the W_i 's are noises. I corresponds to your i -th experiment. So this is the most common situation that you encounter in the lab. If you are dealing with some process, you're trying to measure something you measure it over and over. Each time your measurement has some random error. And then you need to take all your measurements together and come up with a single estimate.

So the noises are assumed to be independent of each other, and also to be independent from the value of the true parameter. Without loss of generality we can assume that the noises have 0 mean and they have some variances that we assume to be known. Theta itself has a prior distribution with a certain mean and the certain variance.

So the form of the optimal linear estimator is really nice. Well maybe you cannot see it right away because this looks messy, but what is it really? It's a linear combination of the X 's and the prior mean. And it's actually a weighted average of the X 's and the prior mean. Here we collect all of the coefficients that we have at the top.

So the whole thing is basically a weighted average. $1/(\sigma_i^2)$ is the weight that we give to X_i , and in the denominator we have the sum of all of the weights. So in the end we're dealing with a weighted average. If μ was equal to 1, and all the X_i 's were equal to 1 then our estimate would also be equal to 1.

Now the form of the weights that we have is interesting. Any given data point is weighted inversely proportional to the variance. What does that say? If my i -th data point has a lot of variance, if W_i is very noisy then X_i is not very useful, is not very reliable. So I'm giving it a small weight. Large variance, a lot of error in my X_i means that I should give it a smaller weight.

If two data points have the same variance, they're of comparable quality, then I'm going to give them equal weight. The other interesting thing is that the prior mean is treated the same way as the X 's. So it's treated as an additional observation. So we're taking a weighted average of the prior mean and of the measurements that we are making. The formula looks as if the prior mean was just another data point. So that's the way of thinking about Bayesian estimation.

You have your real data points, the X 's that you observe, you also had some prior information. This plays a role similar to a data point. Interesting note that if all random variables are normal in this model these optimal linear estimator happens to be also the conditional expectation. That's the nice thing about normal random variables that conditional expectations turn out to be linear.

So the optimal estimate and the optimal linear estimate turn out to be the same. And that gives us another interpretation of linear estimation. Linear estimation is essentially the same as pretending that all random variables are normal. So that's a side point. Now I'd like to close with a comment.

You do your measurements and you estimate Θ on the basis of X . Suppose that instead you have a measuring device that's measures X^3 instead of measuring X , and you want to estimate Θ . Are you going to get to different a estimate? Well X and X^3 contained the same information. Telling you X is the same as telling you the value of X^3 .

So the posterior distribution of Theta given X is the same as the posterior distribution of Theta given X-cubed. And so the means of these posterior distributions are going to be the same. So doing transformations through your data does not matter if you're doing optimal least squares estimation. On the other hand, if you restrict yourself to doing linear estimation then using a linear function of X is not the same as using a linear function of X-cubed. So this is a linear estimator, but where the data are the X-cube's, and we have a linear function of the data.

So this means that when you're using linear estimation you have some choices to make linear on what? Sometimes you want to plot your data on a not ordinary scale and try to plot a line through them. Sometimes you plot your data on a logarithmic scale, and try to plot a line through them. Which scale is the appropriate one? Here it would be a cubic scale. And you have to think about your particular model to decide which version would be a more appropriate one.

Finally when we have multiple data sometimes these multiple data might contain the same information. So X is one data point, X-squared is another data point, X-cubed is another data point. The three of them contain the same information, but you can try to form a linear function of them. And then you obtain a linear estimator that has a more general form as a function of X.

So if you want to estimate your Theta as a cubic function of X, for example, you can set up a linear estimation model of this particular form and find the optimal coefficients, the a's and the b's.

All right, so the last slide just gives you the big picture of what's happening in Bayesian Inference, it's for you to ponder. Basically we talked about three possible estimation methods. Maximum posteriori, mean squared error estimation, and linear mean squared error estimation, or least squares estimation. And there's a number of standard examples that you will be seeing over and over in the recitations, tutorial, homework, and so on, perhaps on exams even. Where we take some nice priors on some unknown parameter, we take some nice models for the noise or the observations, and then you need to work out posterior distributions in the various estimates and compare them.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Problem Set 10
Due December 2, 2010 (in recitation)

1. **A financial parable.** An investment bank is managing \$1 billion, which it invests in various financial instruments (“assets”) related to the housing market (e.g., the infamous “mortgage backed securities”). Because the bank is investing with borrowed money, its actual assets are only \$50 million (5%). Accordingly, if the bank loses more than 5%, it becomes insolvent. (Which means that it will have to be bailed out, and the bankers may need to forgo any huge bonuses for a few months.)
 - (a) The bank considers investing in a single asset, whose gain (over a 1-year period, and measured in percentage points) is modeled as a normal random variable R , with mean 7 and standard deviation 10. (That is, the asset is expected to yield a 7% profit.) What is the probability that the bank will become insolvent? Would you accept this level of risk?
 - (b) In order to safeguard its position, the bank decides to diversify its investments. It considers investing \$50 million in each of twenty different assets, with the i th one having a gain R_i , which is again normal with mean 7 and standard deviation 10; the bank’s gain will be $(R_1 + \dots + R_{20})/20$. These twenty assets are chosen to reflect the housing sectors at different states or even countries, and the bank’s rocket scientists choose to model the R_i as independent random variables. According to this model, what is the probability that the bank becomes insolvent?
 - (c) Based on the calculations in part (b), the bank goes ahead with the diversified investment strategy. It turns out that a global economic phenomenon can affect the housing sectors in different states and countries simultaneously, and therefore the gains R_i are in fact positively correlated. Suppose that for every i and j where $i \neq j$, the correlation coefficient $\rho(R_i, R_j)$ is equal to 1/2. What is the probability that the bank becomes insolvent? You can assume that $(R_1 + \dots + R_{20})/20$ is normal.
2. The adult population of Nowhereville consists of 300 males and 196 females. Each male (respectively, female) has a probability of 0.4 (respectively, 0.5) of casting a vote in the local elections, independently of everyone else. Find a good numerical approximation for the probability that more males than females cast a vote.
3. Let S_n be the number of successes in n independent Bernoulli trials, where the probability of success in each trial is $p = \frac{1}{2}$. Provide a numerical value for the limit as n tends to infinity for each of the following three expressions:
 - (a) $\mathbf{P}(\frac{n}{2} - 10 \leq S_n \leq \frac{n}{2} + 10)$
 - (b) $\mathbf{P}(\frac{n}{2} - \frac{n}{10} \leq S_n \leq \frac{n}{2} + \frac{n}{10})$
 - (c) $\mathbf{P}(\frac{n}{2} - \frac{\sqrt{n}}{2} \leq S_n \leq \frac{n}{2} + \frac{\sqrt{n}}{2})$
4. Alice has two coins. The probability of heads for the first coin is 1/3; the probability of heads for the second coin is 2/3. Other than this difference in their bias, the coins are indistinguishable through any measurement known to man. Alice chooses one of the coins randomly and sends it to Bob. Let p be the probability that Alice chose the first coin. Bob tries to guess which of the two coins he received by flipping it 3 times in a row and observing the outcome. Assume that all coin flips are independent. Let Y be the number of heads Bob observed.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
 (Fall 2010)

- (a) Given that Bob observed k heads, what is the probability that he received the first coin?
 - (b) Find values of k for which the probability that Alice sent the first coin increases after Bob observes k heads out of 3 tosses. In other words, for what values of k is the probability that Alice sent the first coin given that Bob observed k heads greater than p ? If we increase p , how does your answer change (goes up, goes down, or stays unchanged)?
 - (c) Help Bob develop the rule for deciding which coin he received based on the number of heads k he observed in 3 tosses if his goal is to minimize the probability of error.
 - (d) For this part, assume $p = 2/3$.
 - i. Find the probability that Bob will guess the coin correctly using the rule above.
 - ii. How does this compare to the probability of guessing correctly if Bob tried to guess which coin he received before flipping it?
 - (e) If we increase p , how does that affect the decision rule?
 - (f) Find the values of p for which Bob will never guess he received the first coin, regardless of the outcome of the tosses.
 - (g) Find the values of p for which Bob will always guess he received the first coin, regardless of the outcome of the tosses.
5. Consider a Bernoulli process X_1, X_2, X_3, \dots with unknown probability of success q . As usual, define the k th inter-arrival time T_k as

$$T_1 = Y_1, \quad T_k = Y_k - Y_{k-1}, \quad k = 2, 3, \dots$$

where Y_k is the time of the k th success. This problem explores estimation of q from observed inter-arrival times $\{t_1, t_2, t_3, \dots\}$.

You may find the following integral useful: For any non-negative integers k and m ,

$$\int_0^1 q^k (1-q)^m dq = \frac{k! m!}{(k+m+1)!}$$

Assume q is sampled from the random variable Q which is uniformly distributed over $[0, 1]$.

- (a) Compute the PMF of T_1 , $p_{T_1}(t_1)$
- (b) Compute the least squares estimate (LSE) of Q from the first recording $T_1 = t_1$.
- (c) Compute the maximum a posteriori (MAP) estimate of Q given the k recordings, $T_1 = t_1, \dots, T_k = t_k$.

For this part only assume q is sampled from the random variable Q which is now uniformly distributed over $[0.5, 1]$

- (d) Find the linear least squares estimate (LLSE) of the second inter-arrival time (T_2), from the observed first arrival time ($T_1 = t_1$).

6. The joint PDF of X and Y is defined as follows:

$$f_{X,Y}(x, y) = \begin{cases} cxy & \text{if } 0 < x \leq 1, 0 < y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Fall 2010)

- (a) Find the normalization constant c .
- (b) Compute the conditional expectation estimator of X based on the observed value $Y = y$.
- (c) Is this estimate different from what you would have guessed before you saw the value $Y = y$? Explain.
- (d) Repeat (b) and (c) for the MAP estimator.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Problem Set 10 Solutions

1. A financial parable.

- (a) The bank becomes insolvent if the asset's gain $R \leq -5$ (i.e., it loses more than 5%). This probability is the CDF of R evaluated at -5 . Since R is normally distributed, we can convert this CDF to be in terms of a standard normal random variable by subtracting away the mean and dividing by the standard deviation, and then look up the value in a standard normal CDF table.

$$\begin{aligned}\mathbf{E}[R] &= 7, \\ \text{var}(R) &= 10^2 = 100, \\ \mathbf{P}(R \leq -5) &= \mathbf{P}\left(\frac{R - 7}{10} \leq \frac{-5 - 7}{10}\right) = \Phi(-1.2) \approx 0.115.\end{aligned}$$

Thus, by investing in just this one asset, the bank has a 11.5% chance of becoming insolvent.

- (b) If we model the R_i 's as **independent** normal random variables, then their sum $R = (R_1 + \dots + R_{20})/20$ is also a normal random variable (see Example 4.11 on page 214 of the text). Thus, we can calculate the mean and variance of this new R and proceed as in part (a). Note that since the random variables are assumed to be independent, the variance of their sum is just the sum of their individual variances.

$$\begin{aligned}\mathbf{E}[R] &= (\mathbf{E}[R_1] + \dots + \mathbf{E}[R_{20}])/20 = 7, \\ \text{var}(R) &= \frac{1}{20^2}(\text{var}(R_1) + \dots + \text{var}(R_{20})) = \frac{20 \cdot 100}{400} = 5, \\ \mathbf{P}(R \leq -5) &= \mathbf{P}\left(\frac{R - 7}{\sqrt{5}} \leq \frac{-5 - 7}{\sqrt{5}}\right) = \Phi(-5.367) \approx 0.0000000439 = 4.39 \cdot 10^{-8}.\end{aligned}$$

Thus, by diversifying and assuming that the 20 assets have **independent** gains, the bank has seemingly decreased its probability of becoming insolvent to a palatable value.

- (c) Now, if the gains R_i are positively correlated, then we can no longer sum up the individual variances; we need to account for the covariance between pairs of random variables. The covariance is given by

$$\text{cov}(R_i, R_j) = \rho(R_i, R_j)\sqrt{\text{var}(R_i)\text{var}(R_j)} = \frac{1}{2}\sqrt{10^2 \cdot 10^2} = 50.$$

From page 220 in the text, we know that the variance in this case is

$$\begin{aligned}\text{var}(R) &= \text{var}\left(\frac{1}{20} \sum_{i=1}^{20} R_i\right) = \frac{1}{400} \left(\sum_{i=1}^{20} \text{var}(R_i) + \sum_{\{(i,j)|i \neq j\}} \text{cov}(R_i, R_j) \right) \\ &= \frac{1}{400} (20 \cdot 100 + 380 \cdot 50) = 52.5.\end{aligned}$$

Since we assume that $R = (R_1 + \dots + R_{20})/20$ is still normal, we can again apply the same steps as in parts (a) and (b):

$$\begin{aligned}\mathbf{E}[R] &= (\mathbf{E}[R_1] + \dots + \mathbf{E}[R_{20}])/20 = 7, \\ \text{var}(R) &= 52.5, \\ \mathbf{P}(R \leq -5) &= \mathbf{P}\left(\frac{R - 7}{\sqrt{52.5}} \leq \frac{-5 - 7}{\sqrt{52.5}}\right) = \Phi(-1.656) \approx 0.0488.\end{aligned}$$

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
 (Fall 2010)

Thus, by taking into account the positive correlation between the assets' gains, we are no longer as comfortable with the probability of insolvency as we thought we were in part (b).

2. Let M and N be the number of males and females, respectively, that cast a vote. We need to find $P(M > N)$, i.e., $P(M - N > 0)$. The central limit theorem does not apply directly to the random variable $M - N$. However, the central limit theorem implies that M and N are well approximated by normal random variables. So, $M - N$ is the difference of two independent approximately normal random variables. Since the difference of two normal random variables is itself normal, it follows that $M - N$ is approximately normal. The mean and variance of $M - N$ are found by

$$\begin{aligned}\mathbf{E}[M - N] &= 300 \cdot 0.4 + 196 \cdot 0.5 = 120 - 98 = 22, \\ \text{var}(M - N) &= \text{var}(M) + \text{var}(N) = 300 \cdot 0.4 \cdot 0.6 + 196 \cdot 0.5 \cdot 0.5 = 121.\end{aligned}$$

Thus, the standard deviation of $M - N$ is 11. Let Z be a standard normal random variable. Using the central limit theorem approximation, we obtain

$$\begin{aligned}\mathbf{P}(M - N > 0) &= \mathbf{P}\left(\frac{M - N - 22}{11} > -\frac{22}{11}\right) \\ &\approx \mathbf{P}(Z \geq -2) \\ &= 0.9772.\end{aligned}$$

A slightly more refined estimate is obtained by expressing the event of interest as $\mathbf{P}(M - N \geq 1/2)$. We then have

$$\begin{aligned}\mathbf{P}(M - N > 1/2) &= \mathbf{P}\left(\frac{M - N - 22}{11} \geq -\frac{21.5}{11}\right) \\ &\approx \mathbf{P}(Z \geq -1.95) \\ &= 0.974.\end{aligned}$$

3. (a) Using the Central Limit Theorem, we obtain $\mathbf{P}(\frac{n}{2} - 10 \leq S_n \leq \frac{n}{2} + 10) \approx \Phi(\frac{20}{\sqrt{n}}) - \Phi(-\frac{20}{\sqrt{n}}) \rightarrow 0$ as $n \rightarrow \infty$.
- (b) The limit is 1, by the weak law of large numbers.
- (c) Using the Central Limit Theorem, we obtain $\mathbf{P}(\frac{n}{2} - \frac{\sqrt{n}}{2} \leq S_n \leq \frac{n}{2} + \frac{\sqrt{n}}{2}) \rightarrow \Phi(1) - \Phi(-1) = 0.6826$.
4. (a) Let C denote the coin that Bob received, so that $C = 1$ if Bob received the first coin, and $C = 2$ if Bob received the second coin. Then $\mathbf{P}(C = 1) = p$ and $\mathbf{P}(C = 2) = 1 - p$. Given C , the number of heads Y in 3 independent tosses is a binomial random variable.
 We can find the probability that Bob received the first coin given that he observed k heads using Bayes' rule.

$$\begin{aligned}
 \mathbf{P}(C = 1 \mid Y = k) &= \frac{\mathbf{P}(Y = k \mid C = 1) \cdot \mathbf{P}(C = 1)}{\mathbf{P}(Y = k \mid C = 1) \cdot \mathbf{P}(C = 1) + \mathbf{P}(Y = k \mid C = 2) \cdot \mathbf{P}(C = 2)} \\
 &= \frac{\binom{3}{k} \cdot (1/3)^k (2/3)^{3-k} p}{\binom{3}{k} \cdot (1/3)^k (2/3)^{3-k} \cdot p + \binom{3}{k} \cdot (2/3)^k (1/3)^{3-k} \cdot (1-p)} \\
 &= \frac{2^{3-k} p}{2^{3-k} p + 2^k (1-p)} = \frac{1}{1 + \frac{1-p}{p} 2^{2k-3}}
 \end{aligned}$$

(b) We want to find k so that the following inequality holds.

$$\begin{aligned}
 \mathbf{P}(C = 1 \mid Y = k) &> p \\
 \frac{2^{3-k} p}{2^{3-k} p + 2^k (1-p)} &> p
 \end{aligned}$$

Note that if $p = 0$ or $p = 1$, there is no value of k that satisfies the inequality. We now solve it for $0 < p < 1$:

$$\begin{aligned}
 \frac{2^{3-k}}{2^{3-k} p + 2^k (1-p)} &> 1 \\
 2^{3-k} &> 2^{3-k} p + 2^k (1-p) \\
 2^{3-k} (1-p) &> 2^k (1-p) \\
 2^{3-k} &> 2^k \\
 2k &< 3 \\
 k &< 3/2
 \end{aligned}$$

For $0 < p < 1$, $k = 0$ or $k = 1$ the probability that Alice sent the first coin increases. The inequality does not depend on p , and so does not change when p increases. Intuitively, this makes sense: lower values of k increase Bob's belief he got the coin with lower probability of heads.

- (c) Given that Bob observes k heads, Bob must decide on whether the first or second coin was used. To minimize the error, he should decide it is the first coin when $\mathbf{P}(C = 1 \mid Y = k) \geq \mathbf{P}(C = 2 \mid Y = k)$. Thus, we have the decision rule given by

$$\begin{aligned}
 \mathbf{P}(C = 1 \mid Y = k) &\geq \mathbf{P}(C = 2 \mid Y = k) \\
 \frac{2^{3-k} p}{2^{3-k} p + 2^k (1-p)} &\geq \frac{2^k (1-p)}{2^{3-k} p + 2^k (1-p)} \\
 2^{3-k} p &\geq 2^k (1-p) \\
 2^{2k-3} &\leq \frac{p}{1-p} \\
 k &\leq \frac{3}{2} + \frac{1}{2} \log_2 \frac{p}{1-p}
 \end{aligned}$$

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
 (Fall 2010)

- (d) i. If $p = 2/3$, the threshold in the rule above is equal to $\frac{3+\log_2 2}{2} = 2$. Therefore, Bob will decide that he received the first coin when he observes 0, 1 or 2 heads, and will decide that he received the second coin when he observes 3 heads.

We find the probability of a correct decision using the total probability law:

$$\begin{aligned}\mathbf{P}(\text{Correct}) &= \mathbf{P}(\text{Correct} \mid C = 1) \cdot p + \mathbf{P}(\text{Correct} \mid C = 2) \cdot (1 - p) \\ &= \mathbf{P}(Y < 3 \mid C = 1) \cdot p + \mathbf{P}(Y = 3 \mid C = 2) \cdot (1 - p) \\ &= (1 - \mathbf{P}(Y = 3 \mid C = 1)) \cdot p + \mathbf{P}(Y = 3 \mid C = 2) \cdot (1 - p) \\ &= (1 - (1/3)^3)(2/3) + (2/3)^3(1/3) = 20/27 \approx .741\end{aligned}$$

- ii. In absence of any data, all Bob can do is decide he received the first coin with some probability q . Note that this rule includes the deterministic decisions that he received either the first coin ($q = 1$) or the second coin ($q = 0$).

In this case, the probability of correct decision is equal to

$$\begin{aligned}\mathbf{P}(\text{Correct}) &= \mathbf{P}(\text{Correct} \mid C = 1) \cdot p + \mathbf{P}(\text{Correct} \mid C = 2) \cdot (1 - p) \\ &= qp + (1 - q)(1 - p) = 1 - p + q(2p - 1) = \frac{1+q}{3}\end{aligned}$$

Clearly, the probability of the correct decision is maximized (or the probability of error is minimized) when $q = 1$, i.e., when Bob deterministically decides he received the first coin. In this case, $\mathbf{P}(\text{Correct}) = 2/3 \approx .667$. Observing 3 coin tosses increases the probability of the correct decision by $2/27 \approx .074$.

- (e) If p is increased, the threshold in the decision rule in part (c) goes up, i.e., the range of values of k for which Bob decides he received the first coin can only go up.
 (f) Bob will never decide he received the first coin if the threshold in the rule above is below zero:

$$\begin{aligned}\frac{3}{2} + \frac{1}{2} \log_2 \frac{p}{1-p} &< 0 \\ \log_2 \frac{p}{1-p} &< -3 \\ \frac{p}{1-p} &< \frac{1}{8} \\ p &< \frac{1}{9}\end{aligned}$$

If $p < 1/9$, the prior probability of receiving the first coin is so low that no amount of evidence from 3 tosses of the coin will make Bob decide he received the first coin.

- (g) Bob will always decide he received the first coin if the threshold in the rule above is equal to or above 3:

$$\begin{aligned}\frac{3}{2} + \frac{1}{2} \log_2 \frac{p}{1-p} &\geq 3 \\ \log_2 \frac{p}{1-p} &\geq 3 \\ \frac{p}{1-p} &\geq 8 \\ p &\geq \frac{8}{9}\end{aligned}$$

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
 (Fall 2010)

If $p \geq 8/9$, the prior probability of receiving the first coin is so high that no amount of evidence from 3 tosses of the coin will make Bob decide he received the second coin.

5. (a) Using the total probability theorem, we have

$$p_{T_1}(t) = \int_0^1 p_{T_1|Q}(t, q) f_Q(q) dq = \int_0^1 (1-q)^{t-1} q dq = \frac{1}{(t+1)t} \quad \text{for } t = 1, 2, \dots$$

- (b) The least squares estimate coincides with the conditional expectation of Q given T_1 , which is derived as

$$\begin{aligned} \mathbf{E}[Q | T_1 = t] &= \int_0^1 p_{Q|T_1}(q | t) q dq \\ &= \int_0^1 \frac{p_{T_1|Q}(t | q) f_Q(q)}{p_{T_1}(t)} q dq \\ &= \int_0^1 t(t+1)q(1-q)^{t-1} q dq \\ &= \int_0^1 t(t+1)q^2(1-q)^{t-1} dq \\ &= t(t+1) \frac{2(t-1)!}{(t+2)!} \\ &= \frac{2}{t+2} \end{aligned}$$

- (c) We write the posterior probability distribution of Q given $T_1 = t_1, \dots, T_k = t_k$

$$\begin{aligned} f_{Q|T_1, \dots, T_k}(q | t_1, \dots, t_k) &= \frac{f_Q(q) \prod_i^k P_{T_i}(T_i = t_i | Q = q)}{\int_0^1 f_Q(q) \prod_i^k P_{T_i}(T_i = t_i | Q = q) dq} \\ &= \frac{q^k (1-q)^{\sum_i^k t_i - k}}{c} \\ &= \frac{1}{c} q^k (1-q)^{\sum_i^k t_i - k}, \end{aligned}$$

where the denominator integrates out q so it could be viewed as a constant scalar c .

To maximize the above probability we set its derivative with respect to q to zero

$$kq^{k-1}(1-q)^{\sum_i^k t_i - k} - (\sum_i^k t_i - k)q^k(1-q)^{\sum_i^k t_i - k - 1} = 0,$$

or equivalently

$$k(1-q) - (\sum_i^k t_i - k)q = 0,$$

which yields the MAP estimate

$$\hat{q} = \frac{k}{\sum_{i=1}^k t_i}.$$

For this part only assume q is sampled from the random variable Q which is now uniformly distributed over $[0.5, 1]$

(d) The LLSE of T_1 given T_2 is

$$\hat{T}_2 = \mathbf{E}[T_2] + \frac{\text{cov}(T_1, T_2)}{\text{var}(T_1)}(T_1 - \mathbf{E}[T_1]),$$

where the coefficients are

$$\mathbf{E}[T_1] = \mathbf{E}[T_2] = \int_{0.5}^1 f_Q(q) \mathbf{E}[T|Q=q] dq = \int_{0.5}^1 2 * 1/q dq = 2 \ln 2,$$

and from the law of total variance

$$\begin{aligned} \text{var}(T_1) &= \text{var}(T_2) = \mathbf{E}[\text{var}(T_1 | Q)] + \text{var}[\mathbf{E}(T_1 | Q)] \\ &= \mathbf{E}\left[\frac{1-Q}{Q^2}\right] + \text{var}\left[\frac{1}{Q}\right] \\ &= \mathbf{E}[1/Q^2] - \mathbf{E}[1/Q] + \mathbf{E}[1/Q^2] - \mathbf{E}[1/Q]^2 \\ &= \int_{0.5}^2 f_Q(q) \frac{1}{q^2} dq - \int_{0.5}^2 f_Q(q) \frac{1}{q} dq + \int_{0.5}^2 f_Q(q) \frac{1}{q^2} dq - \left(\int_{0.5}^2 f_Q(q) \frac{1}{q} dq\right)^2 \\ &= 2 - 2 \ln 2 + 2 - (2 \ln 2)^2 \\ &= 4 - 2 \ln 2 - (2 \ln 2)^2, \end{aligned}$$

and their covariance

$$\begin{aligned} \text{cov}(T_1, T_2) &= \mathbf{E}[T_1 T_2] - \mathbf{E}[T_1] \mathbf{E}[T_2] \\ &= \mathbf{E}[\mathbf{E}[T_1 T_2 | Q]] - \mathbf{E}[T_1] \mathbf{E}[T_2] \\ &= \mathbf{E}[\mathbf{E}[T_1 | Q] \mathbf{E}[T_2 | Q]] - \mathbf{E}[T_1] \mathbf{E}[T_2] \\ &= \mathbf{E}[1/Q^2] - \mathbf{E}[T_1] \mathbf{E}[T_2] \\ &= 2 - 4(\ln 2)^2 \end{aligned}$$

Therefore we have derived the linear least squares estimator

$$\hat{T}_2 = 2 \ln 2 + \frac{2 - 4(\ln 2)^2}{4 - 2 \ln 2 - (2 \ln 2)^2} (T_1 - 2 \ln 2) \approx 1.543 + 0.113 T_1.$$

6. (a) To find the normalization constant c we integrate the joint PDF:

$$\int_0^1 \int_0^1 f_{X,Y}(x, y) dy dx = c \int_0^1 \int_0^1 xy dy dx = c \int_0^1 1/2x dx = c/4.$$

Therefore, $c = 4$.

(b) To construct the conditional expectation estimator, we need to find the conditional probability density.

$$f_{X|Y}(x | y) = \frac{f_{X,Y}(x, y)}{f_Y(y)} = \frac{4xy}{\int_0^1 4xy dx} = \frac{4xy}{2y} = 2x, \quad x \in (0, 1]$$

Thus

$$\hat{x}_{\text{CE}}(y) = \mathbf{E}[X | Y = y] = \int_0^1 x \cdot 2x dx = 2/3.$$

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Fall 2010)

- (c) We first note that the conditional probability does not depend on y . Therefore, X and Y are independent, and whether or not we observe $Y = y$ does not affect the estimate in part (b). Another way to see this is to consider that if we do not observe y , we can compute the marginal $f_X(x) = \int_0^1 4xy dy = 2x$ which is equal to the conditional density, and will therefore produce the same estimate.
- (d) Since X and Y are independent, no estimator can make use of the observed value of Y to estimate X . The MAP estimator for X is equal to 1, regardless of what value y we observe, since the conditional (and the marginal) density is maximized at 1.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

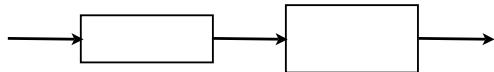
For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

LECTURE 22

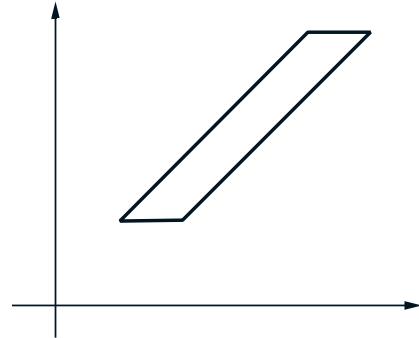
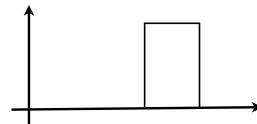
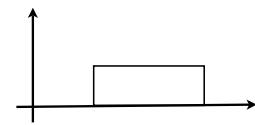
- **Readings:** pp. 225-226; Sections 8.3-8.4

Topics

- (Bayesian) Least means squares (LMS) estimation
- (Bayesian) Linear LMS estimation

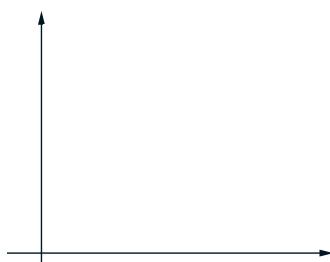
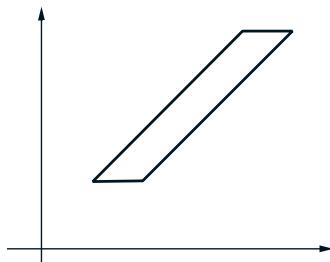


- MAP estimate: $\hat{\theta}_{\text{MAP}}$ maximizes $f_{\Theta|X}(\theta | x)$
- LMS estimation:
 - $\hat{\Theta} = \mathbb{E}[\Theta | X]$ minimizes $\mathbb{E}[(\Theta - g(X))^2]$ over all estimators $g(\cdot)$
 - for any x , $\hat{\theta} = \mathbb{E}[\Theta | X = x]$ minimizes $\mathbb{E}[(\Theta - \hat{\theta})^2 | X = x]$ over all estimates $\hat{\theta}$



Conditional mean squared error

- $E[(\Theta - E[\Theta | X])^2 | X = x]$
 - same as $\text{Var}(\Theta | X = x)$: variance of the conditional distribution of Θ



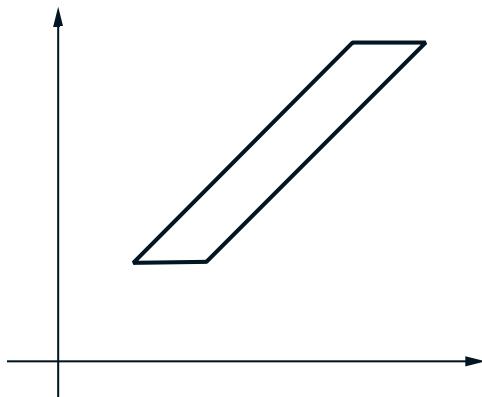
Some properties of LMS estimation

- Estimator: $\hat{\Theta} = \mathbb{E}[\Theta | X]$
- Estimation error: $\tilde{\Theta} = \hat{\Theta} - \Theta$
- $E[\tilde{\Theta}] = 0$ $E[\tilde{\Theta} | X = x] = 0$
- $E[\tilde{\Theta}h(X)] = 0$, for any function h
- $\text{cov}(\tilde{\Theta}, \hat{\Theta}) = 0$
- Since $\Theta = \hat{\Theta} - \tilde{\Theta}$:
 $\text{var}(\Theta) = \text{var}(\hat{\Theta}) + \text{var}(\tilde{\Theta})$

Linear LMS

- Consider estimators of Θ , of the form $\hat{\Theta} = aX + b$
- Minimize $E[(\Theta - aX - b)^2]$
- Best choice of a, b ; best linear estimator:

$$\hat{\Theta}_L = E[\Theta] + \frac{\text{Cov}(X, \Theta)}{\text{var}(X)}(X - E[X])$$



Linear LMS properties

$$\hat{\Theta}_L = E[\Theta] + \frac{\text{Cov}(X, \Theta)}{\text{var}(X)}(X - E[X])$$

$$E[(\hat{\Theta}_L - \Theta)^2] = (1 - \rho^2)\sigma_\Theta^2$$

Linear LMS with multiple data

- Consider estimators of the form:

$$\Theta = a_1X_1 + \cdots + a_nX_n + b$$

- Find best choices of a_1, \dots, a_n, b

- Minimize:

$$E[(a_1X_1 + \cdots + a_nX_n + b - \Theta)^2]$$

- Set derivatives to zero
linear system in b and the a_i

- Only means, variances, covariances matter

The cleanest linear LMS example

$$X_i = \Theta + W_i, \quad \Theta, W_1, \dots, W_n \text{ independent}$$

$$\Theta \sim \mu, \sigma_0^2 \quad W_i \sim 0, \sigma_i^2$$

$$\hat{\Theta}_L = \frac{\mu/\sigma_0^2 + \sum_{i=1}^n X_i/\sigma_i^2}{\sum_{i=0}^n 1/\sigma_i^2}$$

(weighted average of μ, X_1, \dots, X_n)

- If all normal, $\hat{\Theta}_L = E[\Theta | X_1, \dots, X_n]$

Choosing X_i in linear LMS

- $E[\Theta | X]$ is the same as $E[\Theta | X^3]$
- Linear LMS is different:
 - $\hat{\Theta} = aX + b$ versus $\hat{\Theta} = aX^3 + b$
 - Also consider $\hat{\Theta} = a_1X + a_2X^2 + a_3X^3 + b$

Big picture

Standard examples:

- X_i uniform on $[0, \theta]$;
uniform prior on θ
- X_i Bernoulli(p);
uniform (or Beta) prior on p
- X_i normal with mean θ , known variance σ^2 ;
normal prior on θ ;
 $X_i = \Theta + W_i$

Estimation methods:

- MAP
- MSE
- Linear MSE

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013 Transcript – Lecture 23

The following content is provided under a Creative Commons license. Your support will help MIT OpenCourseWare continue to offer high quality, educational resources for free. To make a donation or view additional materials from hundreds of MIT courses, visit MIT OpenCourseWare at ocw.mit.edu.

PROFESSOR: So for the last three lectures we're going to talk about classical statistics, the way statistics can be done if you don't want to assume a prior distribution on the unknown parameters.

Today we're going to focus, mostly, on the estimation side and leave hypothesis testing for the next two lectures. So where there is one generic method that one can use to carry out parameter estimation, that's the maximum likelihood method. We're going to define what it is.

Then we will look at the most common estimation problem there is, which is to estimate the mean of a given distribution. And we're going to talk about confidence intervals, which refers to providing an interval around your estimates, which has some properties of the kind that the parameter is highly likely to be inside that interval, but we will be careful about how to interpret that particular statement.

Ok. So the big framework first. The picture is almost the same as the one that we had in the case of Bayesian statistics. We have some unknown parameter. And we have a measuring device. There is some noise, some randomness.

And we get an observation, X , whose distribution depends on the value of the parameter. However, the big change from the Bayesian setting is that here, this parameter is just a number. It's not modeled as a random variable. It does not have a probability distribution. There's nothing random about it. It's a constant. It just happens that we don't know what that constant is.

And in particular, this probability distribution here, the distribution of X , depends on Θ . But this is not a conditional distribution in the usual sense of the word.

Conditional distributions were defined when we had two random variables and we condition one random variable on the other. And we used the bar to separate the X from the Θ . To make the point that this is not a conditioned distribution, we use a different notation. We put a semicolon here.

And what this is meant to say is that X has a distribution. That distribution has a certain parameter. And we don't know what that parameter is.

So for example, this might be a normal distribution, with variance 1 but a mean Theta. We don't know what Theta is. And we want to estimate it. Now once we have this setting, then your job is to design this box, the estimator.

The estimator is some data processing box that takes the measurements and produces an estimate of the unknown parameter. Now the notation that's used here is as if X and Theta were one-dimensional quantities.

But actually, everything we say remains valid if you interpret X and Theta as vectors of parameters. So for example, you may obtain several measurements, X_1 up to X_n . And there may be several unknown parameters in the background.

Once more, we do not have, and we do not want to assume, a prior distribution on Theta. It's a constant. And if you want to think mathematically about this situation, it's as if you have many different probabilistic models.

So a normal with this mean or a normal with that mean or a normal with that mean, these are alternative candidate probabilistic models. And we want to try to make a decision about which one is the correct model.

In some cases, we have to choose just between a small number of models. For example, you have a coin with an unknown bias. The bias is either 1/2 or 3/4. You're going to flip the coin a few times.

And you try to decide whether the true bias is this one or is that one. So in this case, we have two specific, alternative probabilistic models from which we want to distinguish.

But sometimes things are a little more complicated. For example, you have a coin. And you have one hypothesis that my coin is unbiased. And the other hypothesis is that my coin is biased. And you do your experiments. And you want to come up with a decision that decides whether this is true or this one is true.

In this case, we're not dealing with just two alternative probabilistic models. This one is a specific model for the coin. But this one actually corresponds to lots of possible, alternative coin models.

So this includes the model where Theta is 0.6, the model where Theta is 0.7, Theta is 0.8, and so on. So we're trying to discriminate between one model and lots of alternative models.

How does one go about this? Well, there's some systematic ways that one can approach problems of this kind. And we will start talking about these next time.

So today, we're going to focus on estimation problems. In estimation problems, theta is a quantity, which is a real number, a continuous parameter. We're to design this box, so what we get out of this box is an estimate.

Now notice that this estimate here is a random variable. Even though theta is deterministic, this is random, because it's a function of the data that we observe. The data are random. We're applying a function to the data to construct our estimate.

So, since it's a function of random variables, it's a random variable itself. The distribution of Theta hat depends on the distribution of X. The distribution of X is affected by Theta. So in the end, the distribution of your estimate Theta hat will also be affected by whatever Theta happens to be.

Our general objective, when designing estimators, is that we want to get, in the end, an error, an estimation error, which is not too large. But we'll have to make that specific. Again, what exactly do we mean by that?

So how do we go about this problem? One general approach is to pick a Theta, under which the data that we observe, that this is the X's, our most likely to have occurred.

So I observe X. For any given Theta, I can calculate this quantity, which tells me, under this particular Theta, the X that you observed had this probability of occurring. Under that Theta, the X that you observe had that probability of occurring. You just choose that Theta, which makes the data that you observed most likely.

It's interesting to compare this maximum likelihood estimate with the estimates that you would have, if you were in a Bayesian setting, and you were using maximum approach theory probability estimation.

In the Bayesian setting, what we do is, given the data, we use the prior distribution on Theta. And we calculate the posterior distribution of Theta given X. Notice that this is sort of the opposite from what we have here.

This is the probability of X for a particular value of Theta, whereas this is the probability of Theta for a particular X. So it's the opposite type of conditioning. In the Bayesian setting, Theta is a random variable. So we can talk about the probability distribution of Theta.

So how do these two compare, except for this syntactic difference that the order X's and Theta's are reversed? Let's write down, in full detail, what this posterior distribution of Theta is. By the Bayes rule, this conditional distribution is obtained from the prior, and the model of the measurement process that we have. And we get to this expression.

So in Bayesian estimation, we want to find the most likely value of Theta. And we need to maximize this quantity over all possible Theta's.

First thing to notice is that the denominator is a constant. It does not involve Theta. So when you maximize this quantity, you don't care about the denominator. You just want to maximize the numerator.

Now, here, things start to look a little more similar. And they would be exactly of the same kind, if that term here was absent, if the prior was absent. The two are going to become the same if that prior was just a constant.

So if that prior is a constant, then maximum likelihood estimation takes exactly the same form as Bayesian maximum posterior probability estimation. So you can give this particular interpretation of maximum likelihood estimation.

Maximum likelihood estimation is essentially what you have done, if you were in a Bayesian world, and you had assumed a prior on the Theta's that's uniform, all the Theta's being equally likely.

Okay. So let's look at a simple example. Suppose that the X_i 's are independent, identically distributed random variables, with a certain parameter Theta. So the distribution of each one of the X_i 's is this particular term.

So Theta is one-dimensional. It's a one-dimensional parameter. But we have several data. We write down the formula for the probability of a particular X vector, given a particular value of Theta. But again, when I use the word, given, here it's not in the conditioning sense. It's the value of the density for a particular choice of Theta.

Here, I wrote down, I defined maximum likelihood estimation in terms of PMFs. That's what you would do if the X 's were discrete random variables.

Here, the X 's are continuous random variables, so instead of I'm using the PDF instead of the PMF. So this a definition, here, generalizes to the case of continuous random variables. And you use F 's instead of X 's, our usual recipe.

So the maximum likelihood estimate is defined. Now, since the X_i 's are independent, the joint density of all the X 's together is the product of the individual densities. So you look at this quantity. This is the density or sort of probability of observing a particular sequence of X 's.

And we ask the question, what's the value of Theta that makes the X 's that we observe most likely? So we want to carry out this maximization. Now this maximization is just a calculational problem.

We're going to do this maximization by taking the logarithm of this expression. Maximizing an expression is the same as maximizing the logarithm. So the logarithm of this expression, the logarithm of a product is the sum of the logarithms. You get contributions from this Theta term. There's n of these, so we get an $n \log \Theta$.

And then we have the sum of the logarithms of these terms. It gives us minus Theta. And then the sum of the X 's. So we need to maximize this expression with respect to Theta.

The way to do this maximization is you take the derivative, with respect to Theta. And you get n over Theta equals to the sum of the X 's. And then you solve for Theta. And you find that the maximum likelihood estimate is this quantity.

Which sort of makes sense, because this is the reciprocal of the sample mean of X 's. Theta, in an exponential distribution, we know that it's 1 over (the mean of the exponential distribution). So it looks like a reasonable estimate.

So in any case, this is the estimates that the maximum likelihood estimation procedure tells us that we should report. This formula here, of course, tells you what to do if you have already observed specific numbers. If you have observed specific numbers, then you observe this particular number as your estimate of Theta.

If you want to describe your estimation procedure more abstractly, what you have constructed is an estimator, which is a box that's takes in the random variables, capital X_1 up to Capital X_n , and produces out your estimate, which is also a random variable. Because it's a function of these random variables and is denoted by an upper case Theta, to indicate that this is now a random variable.

So this is an equality about numbers. This is a description of the general procedure, which is an equality between two random variables. And this gives you the more abstract view of what we're doing here.

All right. So what can we tell about our estimate? Is it good or is it bad? So we should look at this particular random variable and talk about the statistical properties that it has.

What we would like is this random variable to be close to the true value of Theta, with high probability, no matter what Theta is, since we don't know what Theta is.

Let's make a little more specific the properties that we want. So we cook up the estimator somehow. So this estimator corresponds, again, to a box that takes data in, the capital X 's, and produces an estimate Theta hat.

This estimate is random. Sometimes it will be above the true value of Theta. Sometimes it will be below. Ideally, we would like it to not have a systematic error, on the positive side or the negative side. So a reasonable wish to have, for a good estimator, is that, on the average, it gives you the correct value.

Now here, let's be a little more specific about what that expectation is. This is an expectation, with respect to the probability distribution of Theta hat. The probability distribution of Theta hat is affected by the probability distribution of the X 's. Because Theta hat is a function of the X 's.

And the probability distribution of the X 's is affected by the true value of Theta. So depending on which one is the true value of Theta, this is going to be a different expectation. So if you were to write this expectation out in more detail, it would look something like this.

You need to write down the probability distribution of Θ . And this is going to be some function. But this function depends on the true Θ , is affected by the true Θ . And then you integrate this with respect to Θ .

What's the point here? Again, Θ is a function of the X 's. So the density of Θ is affected by the density of the X 's. The density of the X 's is affected by the true value of Θ . So the distribution of Θ is affected by the value of Θ .

Another way to put it is, as I've mentioned a few minutes ago, in this business, it's as if we are considering different possible probabilistic models, one probabilistic model for each choice of Θ . And we're trying to guess which one of these probabilistic models is the true one.

One way of emphasizing the fact that this expression depends on the true Θ is to put a little subscript here, expectation, under the particular value of the parameter Θ . So depending on what value the true parameter Θ takes, this expectation will have a different value.

And what we would like is that no matter what the true value is, that our estimate will not have a bias on the positive or the negative sides. So this is a property that's desirable.

Is it always going to be true? Not necessarily, it depends on what estimator we construct. Is it true for our exponential example? Unfortunately not, the estimate that we have in the exponential example turns out to be biased.

And one extreme way of seeing this is to consider the case where our sample size is 1. We're trying to estimate Θ . And the estimator from the previous slide, in that case, is just $1/X_1$. Now X_1 has a fair amount of density in the vicinity of 0, which means that $1/X_1$ has significant probability of being very large.

And if you do the calculation, this ultimately makes the expected value of $1/X_1$ to be infinite. Now infinity is definitely not the correct value. So our estimate is biased upwards. And it's actually biased a lot upwards.

So that's how things are. Maximum likelihood estimates, in general, will be biased. But under some conditions, they will turn out to be asymptotically unbiased.

That is, as you get more and more data, as your X vector is longer and longer, with independent data, the estimate that you're going to have, the expected value of your estimator is going to get closer and closer to the true value. So you do have some nice asymptotic properties, but we're not going to prove anything like this.

Speaking of asymptotic properties, in general, what we would like to have is that, as you collect more and more data, you get the correct answer, in some sense. And the sense that we're going to use here is the limiting sense of convergence in probability, since this is the only notion of convergence of random variables that we have in our hands.

This is similar to what we had in the pollster problem, for example. If we had a bigger and bigger sample size, we could be more and more confident that the estimate that we obtained is close to the unknown true parameter of the distribution that we have.

So this is a desirable property. If you have an infinitely large amount of data, you should be able to estimate an unknown parameter more or less exactly. So this is it desirable property of estimators.

It turns out that maximum likelihood estimation, given independent data, does have this property, under mild conditions. So maximum likelihood estimation, in this respect, is a good approach.

So let's see, do we have this consistency property in our exponential example? In our exponential example, we used this quantity to estimate the unknown parameter Theta. What properties does this quantity have as n goes to infinity?

Well this quantity is the reciprocal of that quantity up here, which is the sample mean. We know from the weak law of large numbers, that the sample mean converges to the expectation. So this property here comes from the weak law of large numbers.

In probability, this quantity converges to the expected value, which, for exponential distributions, is $1/\Theta$. Now, if something converges to something, then the reciprocal of that should converge to the reciprocal of that. That's a property that's certainly correct for numbers.

But you're not talking about convergence of numbers. We're talking about convergence in probability, which is a more complicated notion.

Fortunately, it turns out that the same thing is true, when we deal with convergence in probability. One can show, although we will not bother doing this, that indeed, the reciprocal of this, which is our estimate, converges in probability to the reciprocal of that. And that reciprocal is the true parameter Theta.

So for this particular exponential example, we do have the desirable property, that as the number of data becomes larger and larger, the estimate that we have constructed will get closer and closer to the true parameter value.

And this is true no matter what Theta is. No matter what the true parameter Theta is, we're going to get close to it as we collect more data.

Okay. So these are two rough qualitative properties that would be nice to have. If you want to get a little more quantitative, you can start looking at the mean squared error that your estimator gives.

Now, once more, the comment I was making up there applies. Namely, that this expectation here is an expectation with respect to the probability distribution of Theta hat that corresponds to a particular value of little theta.

So fix a little theta. Write down this expression. Look at the probability distribution of Theta hat, under that little theta. And do this calculation. You're going to get some quantity that depends on the little theta.

And so all quantities in this equality here should be interpreted as quantities under that particular value of little theta. So if you wanted to make this more explicit, you could start throwing little subscripts everywhere in those expressions.

And let's see what those expressions tell us. The expected value squared of a random variable, we know that it's always equal to the variance of this random variable, plus the expectation of the random variable squared. So the expectation value of that random variable, squared.

This equality here is just our familiar formula, that the expected value of X squared is the variance of X plus the expected value of X squared. So we apply this formula to X equal to Theta hat minus Theta.

Now, remember that, in this classical setting, theta is just a constant. We have fixed Theta. We want to calculate the variance of this quantity, under that particular Theta. When you add or subtract a constant to a random variable, the variance doesn't change. This is the same as the variance of our estimator.

And what we've got here is the bias of our estimate. It tells us, on the average, whether we fall above or below. And we're taking the bias to be b squared. If we have an unbiased estimator, the bias term will be 0.

So ideally we want Theta hat to be very close to Theta. And since Theta is a constant, if that happens, the variance of Theta hat would be very small. So Theta is a constant. If Theta hat has a distribution that's concentrated just around own little theta, then Theta hat would have a small variance.

So this is one desire that have. We're going to have a small variance. But we also want to have a small bias at the same time.

So the general form of the mean squared error has two contributions. One is the variance of our estimator. The other is the bias. And one usually wants to design an estimator that simultaneously keeps both of these terms small.

So here's an estimation method that would do very well with respect to this term, but badly with respect to that term. So suppose that my distribution is, let's say, normal with an unknown mean Theta and variance 1.

And I use as my estimator something very dumb. I always produce an estimate that says my estimate is 100. So I'm just ignoring the data and report 100. What does this do?

The variance of my estimator is 0. There's no randomness in the estimate that I report. But the bias is going to be pretty bad. The bias is going to be Θ hat, which is 100 minus the true value of Θ .

And for some Θ 's, my bias is going to be horrible. If my true Θ happens to be 0, my bias squared is a huge term. And I get a large error.

So what's the moral of this example? There are ways of making that variance very small, but, in those cases, you pay a price in the bias. So you want to do something a little more delicate, where you try to keep both terms small at the same time.

So these types of considerations become important when you start to try to design sophisticated estimators for more complicated problems. But we will not do this in this class. This belongs to further classes on statistics and inference.

For this class, for parameter estimation, we will basically stick to two very simple methods. One is the maximum likelihood method we've just discussed. And the other method is what you would do if you were still in high school and didn't know any probability.

You get data. And these data come from some distribution with an unknown mean. And you want to estimate that the unknown mean. What would you do? You would just take those data and average them out.

So let's make this a little more specific. We have X 's that come from a given distribution. We know the general form of the distribution, perhaps. We do know, perhaps, the variance of that distribution, or, perhaps, we don't know it. But we do not know the mean.

And we want to estimate the mean of that distribution. Now, we can write this situation. We can represent it in a different form. The X_i 's are equal to Θ . This is the mean. Plus a 0 mean random variable, that you can think of as noise.

So this corresponds to the usual situation you would have in a lab, where you go and try to measure an unknown quantity. You get lots of measurements. But each time that you measure them, your measurements have some extra noise in there. And you want to kind of get rid of that noise.

The way to try to get rid of the measurement noise is to collect lots of data and average them out. This is the sample mean. And this is a very, very reasonable way of trying to estimate the unknown mean of the X 's.

So this is the sample mean. It's a reasonable, plausible, in general, pretty good estimator of the unknown mean of a certain distribution. We can apply this estimator without really knowing a lot about the distribution of the X 's.

Actually, we don't need to know anything about the distribution. We can still apply it, because the variance, for example, does not show up here. We don't need to know the variance to calculate that quantity.

Does this estimator have good properties? Yes, it does. What's the expected value of the sample mean? If the expectation of this, it's the expectation of this sum divided by n . The expected value for each one of the X 's is Θ . So the expected value of the sample mean is just Θ itself.

So our estimator is unbiased. No matter what Θ is, our estimator does not have a systematic error in either direction. Furthermore, the weak law of large numbers tells us that this quantity converges to the true parameter in probability. So it's a consistent estimator. This is good.

And if you want to calculate the mean squared error corresponding to this estimator. Remember how we defined the mean squared error? It's this quantity. Then it's a calculation that we have done a fair number of times by now.

The mean squared error is the variance of the distribution of the X 's divided by n . So as we get more and more data, the mean squared error goes down to 0.

In some examples, it turns out that the sample mean is also the same as the maximum likelihood estimate. For example, if the X 's are coming from a normal distribution, you can write down the likelihood, do the maximization with respect to Θ , you'll find that the maximum likelihood estimate is the same as the sample mean.

In other cases, the sample mean will be different from the maximum likelihood. And then you have a choice about which one of the two you would use. Probably, in most reasonable situations, you would just use the sample mean, because it's simple, easy to compute, and has nice properties.

All right. So you go to your boss. And you report and say, OK, I did all my experiments in the lab. And the average value that I got is a certain number, 2.37. So is that informative to your boss?

Well your boss would like to know how much they can trust this number, 2.37. Well, I know that the true value is not going to be exactly that. But how close should it be? So give me a range of what you think are possible values of Θ .

So the situation is like this. So suppose that we observe X 's that are coming from a certain distribution. And we're trying to estimate the mean. We get our data. Maybe our data looks something like this.

You calculate the mean. You find the sample mean. So let's suppose that the sample mean is a number, for some reason take to be 2.37. But you want to convey something to your boss about how spread out these data were.

So the boss asks you to give him or her some kind of interval on which Theta, the true parameter, might lie. So the boss asked you for an interval. So what you do is you end up reporting an interval.

And you somehow use the data that you have seen to construct this interval. And you report to your boss also the endpoints of this interval. Let's give names to these endpoints, Θ_{n-} and Θ_{n+} . The ends here just play the role of keeping track of how many data we're using.

So what you report to your boss is this interval as well. Are these Theta's here, the endpoints of the interval, lowercase or uppercase? What should they be? Well you construct these intervals after you see your data.

You take the data into account to construct your interval. So these definitely should depend on the data. And therefore they are random variables. Same thing with your estimator, in general, it's going to be a random variable. Although, when you go and report numbers to your boss, you give the specific realizations of the random variables, given the data that you got.

So instead of having just a single box that produces estimates. So our previous picture was that you have your estimator that takes X 's and produces Θ hats. Now our box will also be producing Θ hats minus and Θ hats plus. It's going to produce an interval as well.

The X 's are random, therefore these quantities are random. Once you go and do the experiment and obtain your data, then your data will be some lowercase x , specific numbers. And then your estimates and estimator become also lower case.

What would we like this interval to do? We would like it to be highly likely to contain the true value of the parameter. So we might impose some specs of the following kind.

I pick a number, alpha. Usually that alpha, think of it as a probability of a large error. Typical value of alpha might be 0.05, in which case this number here is point 0.95.

And you're given specs that say something like this. I would like, with probability at least 0.95, this to happen, which says that the true parameter lies inside the confidence interval.

Now let's try to interpret this statement. Suppose that you did the experiment, and that you ended up reporting to your boss a confidence interval from 1.97 to 2.56. That's what you report to your boss.

And suppose that the confidence interval has this property. Can you go to your boss and say, with probability 95%, the true value of Theta is between these two numbers? Is that a meaningful statement?

So the statement is, the tentative statement is, with probability 95%, the true value of Theta is between 1.97 and 2.56. Well, what is random in that statement? There's nothing random. The true value of theta is a constant. 1.97 is a number. 2.56 is a number.

So it doesn't make any sense to talk about the probability that theta is in this interval. Either theta happens to be in that interval, or it happens to not be. But there are no probabilities associated with this. Because theta is not random.

Syntactically, you can see this. Because theta here is a lower case. So what kind of probabilities are we talking about here? Where's the randomness? Well the random thing is the interval. It's not theta.

So the statement that is being made here is that the interval, that's being constructed by our procedure, should have the property that, with probability 95%, it's going to fall on top of the true value of theta.

So the right way of interpreting what the 95% confidence interval is, is something like the following. We have the true value of theta that we don't know. I get data. Based on the data, I construct a confidence interval. I get my confidence interval. I got lucky. And the true value of theta is in here.

Next day, I do the same experiment, take my data, construct a confidence interval. And I get this confidence interval, lucky once more. Next day I get data. I use my data to come up with an estimate of theta and the confidence interval.

That day, I was unlucky. And I got a confidence interval out there. What the requirement here is, is that 95% of the days, where we use this certain procedure for constructing confidence intervals, 95% of those days, we will be lucky. And we will capture the correct value of theta by your confidence interval.

So it's a statement about the distribution of these random confidence intervals, how likely are they to fall on top of the true theta, as opposed to how likely they are to fall outside. So it's a statement about probabilities associated with a confidence interval. They're not probabilities about theta, because theta, itself, is not random.

So this is what the confidence interval is, in general, and how we interpret it. How do we construct a 95% confidence interval? Let's go through this exercise, in a particular example.

The calculations are exactly the same as the ones that you did when we talked about laws of large numbers and the central limit theorem. So there's nothing new calculationally but it's, perhaps, new in terms of the language that we use and the interpretation.

So we got our sample mean from some distribution. And we would like to calculate a 95% confidence interval. We know from the normal tables, that the standard normal has 2.5% on the tail, that's after 1.96.

Yes, by this time, the number 1.96 should be pretty familiar. So if this probability here is 2.5%, this number here is 1.96.

Now look at this random variable here. This is the sample mean. Difference, from the true mean, normalized by the usual normalizing factor. By the central limit theorem, this is approximately normal. So it has probability 0.95 of being less than 1.96.

Now take this event here and rewrite it. This the event, well, that $\hat{\theta}$ minus θ is bigger than this number and smaller than that number. This event here is equivalent to that event here.

And so this suggests a way of constructing our 95% percent confidence interval. I'm going to report the interval, which gives this as the lower end of the confidence interval, and gives this as the upper end of the confidence interval

In other words, at the end of the experiment, we report the sample mean, which is our estimate. And we report also, an interval around the sample mean. And this is our 95% confidence interval.

The confidence interval becomes smaller, when n is larger. In some sense, we're more certain that we're doing a good estimation job, so we can have a small interval and still be quite confident that our interval captures the true value of the parameter.

Also, if our data have very little noise, when you have more accurate measurements, you're more confident that your estimate is pretty good. And that results in a smaller confidence interval, smaller length of the confidence interval. And still you have 95% probability of capturing the true value of θ .

So we did this exercise by taking 95% confidence intervals and the corresponding value from the normal tables, which is 1.96.

Of course, you can do it more generally, if you set your alpha to be some other number. Again, you look at the normal tables. And you find the value here, so that the tail has probability $\alpha/2$.

And instead of using these 1.96, you use whatever number you get from the normal tables. And this tells you how to construct a confidence interval.

Well, to be exact, this is not necessarily a 95% confidence interval. It's approximately a 95% confidence interval. Why is this? Because we've done an approximation. We have used the central limit theorem.

So it might turn out to be a 95.5% confidence interval instead of 95%, because our calculations are not entirely accurate. But for reasonable values of n , using the central limit theorem is a good approximation. And that's what people almost always do.

So just take the value from the normal tables. Okay, except for one catch. I used the data. I obtained my estimate. And I want to go to my boss and report this $\hat{\theta}$ minus θ , which is the confidence interval.

What's the difficulty? I know what n is. But I don't know what sigma is, in general. So if I don't know sigma, what am I going to do?

Here, there's a few options for what you can do. And the first option is familiar from what we did when we talked about the pollster problem. We don't know what sigma is, but maybe we have an upper bound on sigma.

For example, if the X_i 's Bernoulli random variables, we have seen that the standard deviation is at most $1/2$. So use the most conservative value for sigma. Using the most conservative value means that you take bigger confidence intervals than necessary.

So that's one option. Another option is to try to estimate sigma from the data. How do you do this estimation? In special cases, for special types of distributions, you can think of heuristic ways of doing this estimation.

For example, in the case of Bernoulli random variables, we know that the true value of sigma, the standard deviation of a Bernoulli random variable, is the square root of $\theta(1-\theta)$, where theta is the mean of the Bernoulli.

Try to use this formula. But theta is the thing we're trying to estimate in the first place. We don't know it. What do we do? Well, we have an estimate for theta, the estimate, produced by our estimation procedure, the sample mean.

So I obtain my data. I get my data. I produce the estimate $\hat{\theta}$. It's an estimate of the mean. Use that estimate in this formula to come up with an estimate of my standard deviation. And then use that standard deviation, in the construction of the confidence interval, pretending that this is correct.

Well the number of your data is large, then we know, from the law of large numbers, that $\hat{\theta}$ is a pretty good estimate of theta. So $\hat{\sigma}$ is going to be a pretty good estimate of sigma. So we're not making large errors by using this approach.

So in this scenario here, things were simple, because we had an analytical formula. Sigma was determined by theta. So we could come up with a quick and dirty estimate of sigma.

In general, if you do not have any nice formulas of this kind, what could you do? Well, you still need to come up with an estimate of sigma somehow. What is a generic method for estimating a standard deviation? Equivalently, what could be a generic method for estimating a variance?

Well the variance is an expected value of some random variable. The variance is the mean of the random variable inside of those brackets. How does one estimate the mean of some random variable?

You obtain lots of measurements of that random variable and average them out. So this would be a reasonable way of estimating the variance of a distribution. And again, the weak law of large

numbers tells us that this average converges to the expected value of this, which is just the variance of the distribution.

So we got a nice and consistent way of estimating variances. But now, we seem to be getting in a vicious circle here, because to estimate the variance, we need to know the mean. And the mean is something we're trying to estimate in the first place.

Okay. But we do have an estimate from the mean. So a reasonable approximation, once more, is to plug-in, here, since we don't know the mean, the estimate of the mean. And so you get that expression, but with a theta hat instead of theta itself.

And this is another reasonable way of estimating the variance. It does have the same consistency properties. Why? When n is large, this is going to behave the same as that, because theta hat converges to theta.

And when n is large, this is approximately the same as σ^2 . So for a large n , this quantity also converges to σ^2 . And we have a consistent estimate of the variance as well. And we can take that consistent estimate and use it back in the construction of confidence interval.

One little detail, here, we're dividing by n . Here, we're dividing by $n-1$. Why do we do this? Well, it turns out that's what you need to do for these estimates to be an unbiased estimate of the variance. One has to do a little bit of a calculation, and one finds that that's the factor that you need to have here in order to be unbiased.

Of course, if you get 100 data points, whether you divide by 100 or divided by 99, it's going to make only a tiny difference in your estimate of your variance.

So it's going to make only a tiny difference in your estimate of the standard deviation. It's not a big deal. And it doesn't really matter. But if you want to show off about your deeper knowledge of statistics, you throw in the $1/(n-1)$ factor in there.

So now one basically needs to put together this story here, how you estimate the variance. You first estimate the sample mean. And then you do some extra work to come up with a reasonable estimate of the variance and the standard deviation. And then you use your estimate, of the standard deviation, to come up with a confidence interval, which has these two endpoints.

In doing this procedure, there's basically a number of approximations that are involved. There are two types of approximations. One approximation is that we're pretending that the sample mean has a normal distribution. That's something we're justified to do, by the central limit theorem. But it's not exact. It's an approximation.

And the second approximation that comes in is that, instead of using the correct standard deviation, in general, you will have to use some approximation of the standard deviation.

Okay so you will be getting a little bit of practice with these concepts in recitation and tutorial. And we will move on to new topics next week. But the material that's going to be covered in the final exam is only up to this point. So next week is just general education. Hopefully useful, but it's not in the exam.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

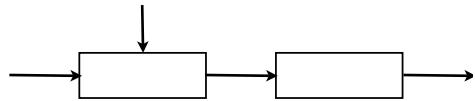
LECTURE 23

- **Readings:** Section 9.1
(not responsible for t -based confidence intervals, in pp. 471-473)

• Outline

- Classical statistics
- Maximum likelihood (ML) estimation
- Estimating a sample mean
- Confidence intervals (CIs)
- CIs using an estimated variance

Classical statistics



- also for vectors X and θ :
 $p_{X_1, \dots, X_n}(x_1, \dots, x_n; \theta_1, \dots, \theta_m)$
- These are NOT conditional probabilities;
 θ is NOT random
 - mathematically: many models, one for each possible value of θ
- **Problem types:**
 - Hypothesis testing:
 $H_0 : \theta = 1/2$ versus $H_1 : \theta = 3/4$
 - Composite hypotheses:
 $H_0 : \theta = 1/2$ versus $H_1 : \theta \neq 1/2$
 - Estimation: design an **estimator** $\hat{\Theta}$, to keep estimation **error** $\hat{\Theta} - \theta$ small

Maximum Likelihood Estimation

- Model, with unknown parameter(s):
 $X \sim p_X(x; \theta)$
- Pick θ that "makes data most likely"

$$\hat{\theta}_{\text{ML}} = \arg \max_{\theta} p_X(x; \theta)$$

- Compare to Bayesian MAP estimation:

$$\hat{\theta}_{\text{MAP}} = \arg \max_{\theta} p_{\Theta|X}(\theta | x)$$

$$\hat{\theta}_{\text{MAP}} = \arg \max_{\theta} \frac{p_{X|\Theta}(x|\theta)p_{\Theta}(\theta)}{p_X(x)}$$

- **Example:** X_1, \dots, X_n : i.i.d., $\text{exponential}(\theta)$

$$\max_{\theta} \prod_{i=1}^n \theta e^{-\theta x_i}$$

$$\max_{\theta} \left(n \log \theta - \theta \sum_{i=1}^n x_i \right)$$

$$\hat{\theta}_{\text{ML}} = \frac{n}{x_1 + \dots + x_n} \quad \hat{\Theta}_n = \frac{n}{X_1 + \dots + X_n}$$

Desirable properties of estimators (should hold FOR ALL θ !!!)

- **Unbiased:** $E[\hat{\Theta}_n] = \theta$
 - exponential example, with $n = 1$:
 $E[1/X_1] = \infty \neq \theta$
(biased)
- **Consistent:** $\hat{\Theta}_n \rightarrow \theta$ (in probability)
 - exponential example:
 $(X_1 + \dots + X_n)/n \rightarrow E[X] = 1/\theta$
 - can use this to show that:
 $\hat{\Theta}_n = n/(X_1 + \dots + X_n) \rightarrow 1/E[X] = \theta$
- **"Small" mean squared error (MSE)**

$$\begin{aligned} E[(\hat{\Theta} - \theta)^2] &= \text{var}(\hat{\Theta} - \theta) + (E[\hat{\Theta} - \theta])^2 \\ &= \text{var}(\hat{\Theta}) + (\text{bias})^2 \end{aligned}$$

Estimate a mean

- X_1, \dots, X_n : i.i.d., mean θ , variance σ^2

$$X_i = \theta + W_i$$

W_i : i.i.d., mean, 0, variance σ^2

$$\hat{\Theta}_n = \text{sample mean} = M_n = \frac{X_1 + \dots + X_n}{n}$$

Properties:

- $E[\hat{\Theta}_n] = \theta$ (unbiased)
- WLLN: $\hat{\Theta}_n \rightarrow \theta$ (consistency)
- MSE: σ^2/n
- Sample mean often turns out to also be the ML estimate.
E.g., if $X_i \sim N(\theta, \sigma^2)$, i.i.d.

Confidence intervals (CIs)

- An estimate $\hat{\Theta}_n$ may not be informative enough

- An $1 - \alpha$ confidence interval is a (random) interval $[\hat{\Theta}_n^- , \hat{\Theta}_n^+]$,

$$\text{s.t. } P(\hat{\Theta}_n^- \leq \theta \leq \hat{\Theta}_n^+) \geq 1 - \alpha, \quad \forall \theta$$

- often $\alpha = 0.05$, or 0.25, or 0.01

- interpretation is subtle

- CI in estimation of the mean

$$\hat{\Theta}_n = (X_1 + \dots + X_n)/n$$

- normal tables: $\Phi(1.96) = 1 - 0.05/2$

$$P\left(\frac{|\hat{\Theta}_n - \theta|}{\sigma/\sqrt{n}} \leq 1.96\right) \approx 0.95 \quad (\text{CLT})$$

$$P\left(\hat{\Theta}_n - \frac{1.96\sigma}{\sqrt{n}} \leq \theta \leq \hat{\Theta}_n + \frac{1.96\sigma}{\sqrt{n}}\right) \approx 0.95$$

More generally: let z be s.t. $\Phi(z) = 1 - \alpha/2$

$$P\left(\hat{\Theta}_n - \frac{z\sigma}{\sqrt{n}} \leq \theta \leq \hat{\Theta}_n + \frac{z\sigma}{\sqrt{n}}\right) \approx 1 - \alpha$$

The case of unknown σ

- Option 1: use upper bound on σ
 - if X_i Bernoulli: $\sigma \leq 1/2$
- Option 2: use ad hoc estimate of σ
 - if X_i Bernoulli(θ): $\hat{\sigma} = \sqrt{\hat{\Theta}(1 - \hat{\Theta})}$
- Option 3: Use generic estimate of the variance
 - Start from $\sigma^2 = E[(X_i - \theta)^2]$

$$\hat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \hat{\Theta}_n)^2 \rightarrow \sigma^2$$

(but do not know θ)

$$\hat{S}_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \hat{\Theta}_n)^2 \rightarrow \sigma^2$$

(unbiased: $E[\hat{S}_n^2] = \sigma^2$)

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013 Transcript – Lecture 24

The following content is provided under a Creative Commons license. Your support will help MIT OpenCourseWare continue to offer high quality educational resources for free. To make a donation or view additional materials from hundreds of MIT courses, visit MIT OpenCourseWare at ocw.mit.edu.

JOHN TSITSIKLIS: And we're going to continue today with our discussion of classical statistics. We'll start with a quick review of what we discussed last time, and then talk about two topics that cover a lot of statistics that are happening in the real world. So two basic methods. One is the method of linear regression, and the other one is the basic methods and tools for how to do hypothesis testing.

OK, so these two are topics that any scientifically literate person should know something about. So we're going to introduce the basic ideas and concepts involved. So in classical statistics we basically have essentially a family of possible models about the world.

So the world is the random variable that we observe, and we have a model for it, but actually not just one model, several candidate models. And each candidate model corresponds to a different value of a parameter theta that we do not know. So in contrast to Bayesian statistics, this theta is assumed to be a constant that we do not know. It is not modeled as a random variable, there's no probabilities associated with theta.

We only have probabilities about the X's. So in this context what is a reasonable way of choosing a value for the parameter? One general approach is the maximum likelihood approach, which chooses the theta for which this quantity is largest. So what does that mean intuitively? I'm trying to find the value of theta under which the data that I observe are most likely to have occurred.

So the thinking is essentially as follows. Let's say I have to choose between two choices of theta. Under this theta the X that I observed would be very unlikely. Under that theta the X that I observed would have a decent probability of occurring. So I chose the latter as my estimate of theta.

It's interesting to do the comparison with the Bayesian approach which we did discuss last time, in the Bayesian approach we also maximize over theta, but we maximize a quantity in which the relation between X's and thetas run the opposite way.

Here in the Bayesian world, Theta is a random variable. So it has a distribution. Once we observe the data, it has a posterior distribution, and we find the value of Theta, which is most likely under the posterior distribution.

As we discussed last time when you do this maximization now the posterior distribution is given by this expression. The denominator doesn't matter, and if you were to take a prior, which is flat-

- that is a constant independent of Theta, then that term would go away. And syntactically, at least, the two approaches look the same.

So syntactically, or formally, maximum likelihood estimation is the same as Bayesian estimation in which you assume a prior which is flat, so that all possible values of Theta are equally likely.

Philosophically, however, they're very different things. Here I'm picking the most likely value of Theta. Here I'm picking the value of Theta under which the observed data would have been more likely to occur. So maximum likelihood estimation is a general purpose method, so it's applied all over the place in many, many different types of estimation problems.

There is a special kind of estimation problem in which you may forget about maximum likelihood estimation, and come up with an estimate in a straightforward way. And this is the case where you're trying to estimate the mean of the distribution of X, where X is a random variable. You observe several independent identically distributed random variables X₁ up to X_n. All of them have the same distribution as this X.

So they have a common mean. We do not know the mean we want to estimate it. What is more natural than just taking the average of the values that we have observed? So you generate lots of X's, take the average of them, and you expect that this is going to be a reasonable estimate of the true mean of that random variable. And indeed we know from the weak law of large numbers that this estimate converges in probability to the true mean of the random variable.

The other thing that we talked about last time is that besides giving a point estimate we may want to also give an interval that tells us something about where we might believe theta to lie. And 1-alpha confidence interval is an interval generated based on the data. So it's an interval from this value to that value. These values are written with capital letters because they're random, because they depend on the data that we have seen. And this gives us an interval, and we would like this interval to have the property that theta is inside that interval with high probability.

So typically we would take 1-alpha to be a quantity such as 95% for example. In which case we have a 95% confidence interval. As we discussed last time it's important to have the right interpretation of what's 95% means.

What it does not mean is the following-- the unknown value has 95% percent probability of being in the interval that we have generated. That's because the unknown value is not a random variable, it's a constant. Once we generate the interval either it's inside or it's outside, but there's no probabilities involved.

Rather the probabilities are to be interpreted over the random interval itself. What a statement like this says is that if I have a procedure for generating 95% confidence intervals, then whenever I use that procedure I'm going to get a random interval, and it's going to have 95% probability of capturing the true value of theta.

So most of the time when I use this particular procedure for generating confidence intervals the true theta will happen to lie inside that confidence interval with probability 95%. So the

randomness in this statement is with respect to my confidence interval, it's not with respect to theta, because theta is not random.

How does one construct confidence intervals? There's various ways of going about it, but in the case where we're dealing with the estimation of the mean of a random variable doing this is straightforward using the central limit theorem. Basically we take our estimated mean, that's the sample mean, and we take a symmetric interval to the left and to the right of the sample mean.

And we choose the width of that interval by looking at the normal tables. So if this quantity, $1 - \alpha$ is 95% percent, we're going to look at the 97.5 percentile of the normal distribution. Find the constant number that corresponds to that value from the normal tables, and construct the confidence intervals according to this formula. So that gives you a pretty mechanical way of going about constructing confidence intervals when you're estimating the sample mean.

So constructing confidence intervals in this way involves an approximation. The approximation is the central limit theorem. We are pretending that the sample mean is a normal random variable. Which is, more or less, right when n is large. That's what the central limit theorem tells us.

And sometimes we may need to do some extra approximation work, because quite often we do not know the true value of σ . So we need to do some work either to estimate σ from the data. So σ is, of course, the standard deviation of the X 's. We may want to estimate it from the data, or we may have an upper bound on σ , and we just use that upper bound.

So now let's move on to a new topic. A lot of statistics in the real world are of the following flavor. So suppose that X is the SAT score of a student in high school, and Y is the MIT GPA of that same student. So you expect that there is a relation between these two. So you go and collect data for different students, and you record for a typical student this would be their SAT score, that could be their MIT GPA. And you plot all this data on an (X, Y) diagram.

Now it's reasonable to believe that there is some systematic relation between the two. So people who had higher SAT scores in high school may have higher GPA in college. Well that may or may not be true. You want to construct a model of this kind, and see to what extent a relation of this type is true.

So you might hypothesize that the real world is described by a model of this kind. That there is a linear relation between the SAT score, and the college GPA. So it's a linear relation with some parameters, θ_0 and θ_1 that we do not know.

So we assume a linear relation for the data, and depending on the choices of θ_0 and θ_1 it could be a different line through those data. Now we would like to find the best model of this kind to explain the data. Of course there's going to be some randomness. So in general it's going to be impossible to find a line that goes through all of the data points.

So let's try to find the best line that comes closest to explaining those data. And here's how we go about it. Suppose we try some particular values of theta₀ and theta₁. These give us a certain line. Given that line, we can make predictions.

For a student who had this x , the model that we have would predict that y would be this value. The actual y is something else, and so this quantity is the error that our model would make in predicting the y of that particular student. We would like to choose a line for which the predictions are as good as possible. And what do we mean by as good as possible? As our criteria we're going to take the following.

We are going to look at the prediction error that our model makes for each particular student. Take the square of that, and then add them up over all of our data points. So what we're looking at is the sum of this quantity squared, that quantity squared, that quantity squared, and so on. We add all of these squares, and we would like to find the line for which the sum of these squared prediction errors are as small as possible.

So that's the procedure. We have our data, the X 's and the Y 's. And we're going to find theta's the best model of this type, the best possible model, by minimizing this sum of squared errors. So that's a method that one could pull out of the hat and say OK, that's how I'm going to build my model. And it sounds pretty reasonable.

And it sounds pretty reasonable even if you don't know anything about probability. But does it have some probabilistic justification? It turns out that yes, you can motivate this method with probabilistic considerations under certain assumptions. So let's make a probabilistic model that's going to lead us to these particular way of estimating the parameters.

So here's a probabilistic model. I pick a student who had a specific SAT score. And that could be done at random, but also could be done in a systematic way. That is, I pick a student who had an SAT of 600, a student of 610 all the way to 1,400 or 1,600, whatever the right number is. I pick all those students.

And I assume that for a student of this kind there's a true model that tells me that their GPA is going to be a random variable, which is something predicted by their SAT score plus some randomness, some random noise. And I model that random noise by independent normal random variables with 0 mean and a certain variance.

So this is a specific probabilistic model, and now I can think about doing maximum likelihood estimation for this particular model. So to do maximum likelihood estimation here I need to write down the likelihood of the y 's that I have observed. What's the likelihood of the y 's that I have observed?

Well, a particular w has a likelihood of the form e to the minus w squared over (2 sigma-squared). That's the likelihood of a particular w . The probability, or the likelihood of observing a particular value of y , that's the same as the likelihood that w takes a value of y minus this, minus that. So the likelihood of the y 's is of this form. Think of this as just being the w_i -squared.

So this is the density -- and if we have multiple data you multiply the likelihoods of the different y 's. So you have to write something like this. Since the w 's are independent that means that the y 's are also independent. The likelihood of a y vector is the product of the likelihoods of the individual y 's. The likelihood of every individual y is of this form. Where w is y_i minus these two quantities.

So this is the form that the likelihood function is going to take under this particular model. And under the maximum likelihood methodology we want to maximize this quantity with respect to θ_0 and θ_1 . Now to do this maximization you might as well consider the logarithm and maximize the logarithm, which is just the exponent up here. Maximizing this exponent because we have a minus sign is the same as minimizing the exponent without the minus sign. σ^2 is a constant. So what you end up doing is minimizing this quantity here, which is the same as what we had in our linear regression methods.

So in conclusion you might choose to do linear regression in this particular way, just because it looks reasonable or plausible. Or you might interpret what you're doing as maximum likelihood estimation, in which you assume a model of this kind where the noise terms are normal random variables with the same distribution -- independent identically distributed.

So linear regression implicitly makes an assumption of this kind. It's doing maximum likelihood estimation as if the world was really described by a model of this form, and with the W 's being random variables. So this gives us at least some justification that this particular approach to fitting lines to data is not so arbitrary, but it has a sound footing.

OK so then once you accept this formulation as being a reasonable one what's the next step? The next step is to see how to carry out this minimization. This is not a very difficult minimization to do. The way it's done is by setting the derivatives of this expression to 0. Now because this is a quadratic function of θ_0 and θ_1 -- when you take the derivatives with respect to θ_0 and θ_1 -- you get linear functions of θ_0 and θ_1 . And you end up solving a system of linear equations in θ_0 and θ_1 . And it turns out that there's very nice and simple formulas for the optimal estimates of the parameters in terms of the data.

And the formulas are these ones. I said that these are nice and simple formulas. Let's see why. How can we interpret them? So suppose that the world is described by a model of this kind, where the X 's and Y 's are random variables. And where W is a noise term that's independent of X . So we're assuming that a linear model is indeed true, but not exactly true. There's always some noise associated with any particular data point that we obtain.

So if a model of this kind is true, and the W 's have 0 mean then we have that the expected value of Y would be θ_0 plus θ_1 expected value of X . And because W has 0 mean there's no extra term. So in particular, θ_0 would be equal to expected value of Y minus θ_1 expected value of X .

So let's use this equation to try to come up with a reasonable estimate of θ_0 . I do not know the expected value of Y , but I can estimate it. How do I estimate it? I look at the average of all

the y's that I have obtained. so I replace this, I estimate it with the average of the data I have seen.

Here, similarly with the X's. I might not know the expected value of X's, but I have data points for the x's. I look at the average of all my data points, I come up with an estimate of this expectation. Now I don't know what theta1 is, but my procedure is going to generate an estimate of theta1 called theta1 hat. And once I have this estimate, then a reasonable person would estimate theta0 in this particular way.

So that's how my estimate of theta0 is going to be constructed. It's this formula here. We have not yet addressed the harder question, which is how to estimate theta1 in the first place. So to estimate theta0 I assumed that I already had an estimate for a theta1.

OK, the right formula for the estimate of theta1 happens to be this one. It looks messy, but let's try to interpret it. What I'm going to do is I'm going to take this model for simplicity let's assume that they're the random variables have 0 means. And see how we might estimate how we might try to estimate theta1.

Let's multiply both sides of this equation by X. So we get Y times X equals theta0 plus theta0 times X plus theta1 times X-squared, plus X times W. And now take expectations of both sides. If I have 0 mean random variables the expected value of Y times X is just the covariance of X with Y.

I have assumed that my random variables have 0 means, so the expectation of this is 0. This one is going to be the variance of X, so I have theta1 times variance of X. And since I'm assuming that my random variables have 0 mean, and I'm also assuming that W is independent of X this last term also has 0 mean.

So under such a probabilistic model this equation is true. If we knew the variance and the covariance then we would know the value of theta1. But we only have data, we do not necessarily know the variance and the covariance, but we can estimate it.

What's a reasonable estimate of the variance? The reasonable estimate of the variance is this quantity here divided by n, and the reasonable estimate of the covariance is that numerator divided by n.

So this is my estimate of the mean. I'm looking at the squared distances from the mean, and I average them over lots and lots of data. This is the most reasonable way of estimating the variance of our distribution.

And similarly the expected value of this quantity is the covariance of X with Y, and then we have lots and lots of data points. This quantity here is going to be a very good estimate of the covariance. So basically what this formula does is-- one way of thinking about it-- is that it starts from this relation which is true exactly, but estimates the covariance and the variance on the basis of the data, and then using these estimates to come up with an estimate of theta1.

So this gives us a probabilistic interpretation of the formulas that we have for the way that the estimates are constructed. If you're willing to assume that this is the true model of the world, the structure of the true model of the world, except that you do not know means and covariances, and variances. Then this is a natural way of estimating those unknown parameters.

All right, so we have a closed-form formula, we can apply it whenever we have data. Now linear regression is a subject on which there are whole courses, and whole books that are given. And the reason for that is that there's a lot more that you can bring into the topic, and many ways that you can elaborate on the simple solution that we got for the case of two parameters and only two random variables.

So let me give you a little bit of flavor of what are the topics that come up when you start looking into linear regression in more depth. So in our discussions so far we made the linear model in which we're trying to explain the values of one variable in terms of the values of another variable. We're trying to explain GPAs in terms of SAT scores, or we're trying to predict GPAs in terms of SAT scores.

But maybe your GPA is affected by several factors. For example maybe your GPA is affected by your SAT score, also the income of your family, the years of education of your grandmother, and many other factors like that. So you might write down a model in which I believe that GPA has a relation, which is a linear function of all these other variables that I mentioned. So perhaps you have a theory of what determines performance at college, and you want to build a model of that type.

How do we go about in this case? Well, again we collect the data points. We look at the i -th student, who has a college GPA. We record their SAT score, their family income, and grandmother's years of education. So this is one data point that is for one particular student.

We postulate the model of this form. For the i -th student this would be the mistake that our model makes if we have chosen specific values for those parameters. And then we go and choose the parameters that are going to give us, again, the smallest possible sum of squared errors. So philosophically it's exactly the same as what we were discussing before, except that now we're including multiple explanatory variables in our model instead of a single explanatory variable.

So that's the formulation. What do you do next? Well, to do this minimization you're going to take derivatives once you have your data, you have a function of these three parameters. You take the derivative with respect to the parameter, set the derivative equal to 0, you get the system of linear equations. You throw that system of linear equations to the computer, and you get numerical values for the optimal parameters.

There are no nice closed-form formulas of the type that we had in the previous slide when you're dealing with multiple variables. Unless you're willing to go into matrix notation. In that case you can again write down closed-form formulas, but they will be a little less intuitive than what we had before. But the moral of the story is that numerically this is a procedure that's very easy. It's a problem, an optimization problem that the computer can solve for you. And it can solve it for you very quickly. Because all that it involves is solving a system of linear equations.

Now when you choose your explanatory variables you may have some choices. One person may think that your GPA has something to do with your SAT score. Some other person may think that your GPA has something to do with the square of your SAT score. And that other person may want to try to build a model of this kind.

Now when would you want to do this? Suppose that the data that you have looks like this. If the data looks like this then you might be tempted to say well a linear model does not look right, but maybe a quadratic model will give me a better fit for the data. So if you want to fit a quadratic model to the data then what you do is you take X^2 as your explanatory variable instead of X , and you build a model of this kind.

There's nothing really different in models of this kind compared to models of that kind. They are still linear models because we have theta's showing up in a linear fashion. What you take as your explanatory variables, whether it's X , whether it's X^2 , or whether it's some other function that you chose. Some general function h of X , doesn't make a difference. So think of you h of X as being your new X . So you can formulate the problem exactly the same way, except that instead of using X 's you choose h of X 's.

So it's basically a question do I want to build a model that explains Y 's based on the values of X , or do I want to build a model that explains Y 's on the basis of the values of h of X . Which is the right value to use? And with this picture here, we see that it can make a difference. A linear model in X might be a poor fit, but a quadratic model might give us a better fit.

So this brings to the topic of how to choose your functions h of X if you're dealing with a real world problem. So in a real world problem you're just given X 's and Y 's. And you have the freedom of building models of any kind you want. You have the freedom of choosing a function h of X of any type that you want.

So this turns out to be a quite difficult and tricky topic. Because you may be tempted to overdo it. For example, I got my 10 data points, and I could say OK, I'm going to choose an h of X . I'm going to choose h of X and actually multiple h 's of X to do a multiple linear regression in which I'm going to build a model that's uses a 10th degree polynomial.

If I choose to fit my data with a 10th degree polynomial I'm going to fit my data perfectly, but I may obtain a model that does something like this, and goes through all my data points. So I can make my prediction errors extremely small if I use lots of parameters, and if I choose my h functions appropriately. But clearly this would be garbage.

If you get those data points, and you say here's my model that explains them. That has a polynomial going up and down, then you're probably doing something wrong. So choosing how complicated those functions, the h 's, should be. And how many explanatory variables to use is a very delicate and deep topic on which there's deep theory that tells you what you should do, and what you shouldn't do.

But the main thing that one should avoid doing is having too many parameters in your model when you have too few data. So if you only have 10 data points, you shouldn't have 10 free

parameters. With 10 free parameters you will be able to fit your data perfectly, but you wouldn't be able to really rely on the results that you are seeing.

OK, now in practice, when people run linear regressions they do not just give point estimates for the parameters theta. But similar to what we did for the case of estimating the mean of a random variable you might want to give confidence intervals that sort of tell you how much randomness there is when you estimate each one of the particular parameters.

There are formulas for building confidence intervals for the estimates of the theta's. We're not going to look at them, it would take too much time. Also you might want to estimate the variance in the noise that you have in your model. That is if you are pretending that your true model is of the kind we were discussing before, namely Y equals $\theta_0 + \theta_1 X + W$, and W has a variance σ^2 . You might want to estimate this, because it tells you something about the model, and this is called standard error.

It puts a limit on how good predictions your model can make. Even if you have the correct θ_0 and θ_1 , and somebody tells you X you can make a prediction about Y , but that prediction will not be accurate. Because there's this additional randomness. And if that additional randomness is big, then your predictions will also have a substantial error in them.

There's another quantity that gets reported usually. This is part of the computer output that you get when you use a statistical package which is called R-square. And its a measure of the explanatory power of the model that you have built linear regression. Using linear regression. Instead of defining R-square exactly, let me give you a sort of analogous quantity that's involved.

After you do your linear regression you can look at the following quantity. You look at the variance of Y , which is something that you can estimate from data. This is how much randomness there is in Y . And compare it with the randomness that you have in Y , but conditioned on X . So this quantity tells me if I knew X how much randomness would there still be in my Y ?

So if I know X , I have more information, so Y is more constrained. There's less randomness in Y . This is the randomness in Y if I don't know anything about X .

So naturally this quantity would be less than 1, and if this quantity is small it would mean that whenever I know X then Y is very well known. Which essentially tells me that knowing X allows me to make very good predictions about Y . Knowing X means that I'm explaining away most of the randomness in Y .

So if you read a statistical study that uses linear regression you might encounter statements of the form 60% of a student's GPA is explained by the family income. If you read the statements of this kind it's really refers to quantities of this kind. Out of the total variance in Y , how much variance is left after we build our model?

So if only 40% of the variance of Y is left after we build our model, that means that X explains 60% of the variations in Y 's. So the idea is that randomness in Y is caused by multiple sources.

Our explanatory variable and random noise. And we ask the question what percentage of the total randomness in Y is explained by variations in the X parameter? And how much of the total randomness in Y is attributed just to random effects? So if you have a model that explains most of the variation in Y then you can think that you have a good model that tells you something useful about the real world.

Now there's lots of things that can go wrong when you use linear regression, and there's many pitfalls. One pitfall happens when you have this situation that's called heteroskedacity. So suppose your data are of this kind. So what's happening here? You seem to have a linear model, but when X is small you have a very good model. So this means that W has a small variance when X is here.

On the other hand, when X is there you have a lot of randomness. This would be a situation in which the W's are not identically distributed, but the variance of the W's, of the noise, has something to do with the X's. So with different regions of our x-space we have different amounts of noise. What will go wrong in this situation? Since we're trying to minimize sum of squared errors, we're really paying attention to the biggest errors. Which will mean that we are going to pay attention to these data points, because that's where the big errors are going to be. So the linear regression formulas will end up building a model based on these data, which are the most noisy ones. Instead of those data that are nicely stacked in order.

Clearly that's not to the right thing to do. So you need to change something, and use the fact that the variance of W changes with the X's, and there are ways of dealing with it. It's something that one needs to be careful about. Another possibility of getting into trouble is if you're using multiple explanatory variables that are very closely related to each other.

So for example, suppose that I tried to predict your GPA by looking at your SAT the first time that you took it plus your SAT the second time that you took your SATs. I'm assuming that almost everyone takes the SAT more than once. So suppose that you had a model of this kind.

Well, SAT on your first try and SAT on your second try are very likely to be fairly close. And you could think of coming up with estimates in which this is ignored. And you build a model based on this, or an alternative model in which this term is ignored, and you make predictions based on the second SAT. And both models are likely to be essentially as good as the other one, because these two quantities are essentially the same.

So in that case, your theta's that you estimate are going to be very sensitive to little details of the data. You change your data, you have your data, and your data tell you that this coefficient is big and that coefficient is small. You change your data just a tiny bit, and your theta's would drastically change. So this is a case in which you have multiple explanatory variables, but they're redundant in the sense that they're very closely related to each other, and perhaps with a linear relation. So one must be careful about the situation, and do special tests to make sure that this doesn't happen.

Finally the biggest and most common blunder is that you run your linear regression, you get your linear model, and then you say oh, OK. Y is caused by X according to this particular formula.

Well, all that we did was to identify a linear relation between X and Y. This doesn't tell us anything. Whether it's Y that causes X, or whether it's X that causes Y, or maybe both X and Y are caused by some other variable that we didn't think about.

So building a good linear model that has small errors does not tell us anything about causal relations between the two variables. It only tells us that there's a close association between the two variables. If you know one you can make predictions about the other. But it doesn't tell you anything about the underlying physics, that there's some physical mechanism that introduces the relation between those variables.

OK, that's it about linear regression. Let us start the next topic, which is hypothesis testing. And we're going to continue with it next time.

So here, instead of trying to estimate continuous parameters, we have two alternative hypotheses about the distribution of the X random variable. So for example our random variable could be either distributed according to this distribution, under H₀, or it might be distributed according to this distribution under H₁. And we want to make a decision which distribution is the correct one?

So we're given those two distributions, and some common terminologies that one of them is the null hypothesis-- sort of the default hypothesis, and we have some alternative hypotheses-- and we want to check whether this one is true, or that one is true. So you obtain a data point, and you want to make a decision. In this picture what would a reasonable person do to make a decision? They would probably choose a certain threshold, X_i , and decide that H₁ is true if your data falls in this interval. And decide that H₀ is true if you fall on the side. So that would be a reasonable way of approaching the problem.

More generally you take the set of all possible X's, and you divide the set of possible X's into two regions. One is the rejection region, in which you decide H₁, or you reject H₀. And the complement of that region is where you decide H₀.

So this is the x-space of your data. In this example here, x was one-dimensional. But in general X is going to be a vector, where all the possible data vectors that you can get, they're divided into two types. If it falls in this set you'd make one decision. If it falls in that set, you make the other decision. OK, so how would you characterize the performance of the particular way of making a decision?

Suppose I chose my threshold. I may make mistakes of two possible types. Perhaps H₀ is true, but my data happens to fall here. In which case I make a mistake, and this would be a false rejection of H₀. If my data falls here I reject H₀. I decide H₁. Whereas H₀ was true. The probability of this happening? Let's call it alpha.

But there's another kind of error that can be made. Suppose that H₁ was true, but by accident my data happens to fall on that side. Then I'm going to make an error again. I'm going to decide H₀ even though H₁ was true. How likely is this to occur? This would be the area under this curve here. And that's the other type of error than can be made, and beta is the probability of this particular type of error.

Both of these are errors. Alpha is the probability of error of one kind. Beta is the probability of an error of the other kind. You would like the probabilities of error to be small. So you would like to make both alpha and beta as small as possible.

Unfortunately that's not possible, there's a trade-off. If I go to my threshold it this way, then alpha becomes smaller, but beta becomes bigger. So there's a trade-off. If I make my rejection region smaller one kind of error is less likely, but the other kind of error becomes more likely. So we got this trade-off.

So what do we do about it? How do we move systematically? How do we come up with rejection regions? Well, what the theory basically tells you is it tells you how you should create those regions. But it doesn't tell you exactly how. It tells you the general shape of those regions.

For example here, the theory who tells us that the right thing to do would be to put the threshold and make decisions one way to the right, one way to the left. But it might not necessarily tell us where to put the threshold. Still, it's useful enough to know that the way to make a good decision would be in terms of a particular threshold.

Let me make this more specific. We can take our inspiration from the solution of the hypothesis testing problem that we had in the Bayesian case. In the Bayesian case we just pick the hypothesis which is more likely given the data. The produced posterior probabilities using Bayesian rule, they're written this way.

And this term is the same as that term. They cancel out, then let me collect terms here and there. I get an expression here. I think the version you have in your handout is the correct one. The one on the slide was not the correct one, so I'm fixing it here.

OK, so this is the form of how you make decisions in the Bayesian case. What you do in the Bayesian case, you calculate this ratio. Let's call it the likelihood ratio. And compare that ratio to a threshold. And the threshold that you should be using in the Bayesian case has something to do with the prior probabilities of the two hypotheses.

In the non-Bayesian case we do not have prior probabilities, so we do not know how to set this threshold. But we're going to do is we're going to keep this particular structure anyway, and maybe use some other considerations to pick the threshold. So we're going to use a likelihood ratio test, that's how it's called in which we calculate a quantity of this kind that we call the likelihood, and compare it with a threshold.

So what's the interpretation of this likelihood? We ask-- the X's that I have observed, how likely were they to occur if H₁ was true? And how likely were they to occur if H₀ was true? This ratio could be big if my data are plausible they might occur under H₁. But they're very implausible, extremely unlikely to occur under H₀.

Then my thinking would be well the data that I saw are extremely unlikely to have occurred under H₀. So H₀ is probably not true. I'm going to go for H₁ and choose H₁. So when this ratio is big it tells us that the data that we're seeing are better explained if we assume H₁ to be true

rather than H_0 to be true. So I calculate this quantity, compare it with a threshold, and that's how I make my decision.

So in this particular picture, for example the way it would go would be the likelihood ratio in this picture goes monotonically with my X . So comparing the likelihood ratio to the threshold would be the same as comparing my x to the threshold, and we've got the question of how to choose the threshold.

The way that the threshold is chosen is usually done by fixing one of the two probabilities of error. That is, I say, that I want my error of one particular type to be a given number, so I fix this α . And then I try to find where my threshold should be. So that this probability θ , probability out there, is just equal to α .

And then the other probability of error, β , will be whatever it turns out to be. So somebody picks α ahead of time. Based on the probability of a false rejection based on α , I find where my threshold is going to be. I choose my threshold, and that determines subsequently the value of β . So we're going to continue with this story next time, and we'll stop here.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

LECTURE 24

- Reference: Section 9.3

Outline

- Review
 - Maximum likelihood estimation
 - Confidence intervals
- Linear regression
- Binary hypothesis testing
 - Types of error
 - Likelihood ratio test (LRT)

Review

- Maximum likelihood estimation
 - Have model with unknown parameters: $X \sim p_X(x; \theta)$
 - Pick θ that “makes data most likely”

$$\max_{\theta} p_X(x; \theta)$$
 - Compare to Bayesian MAP estimation:

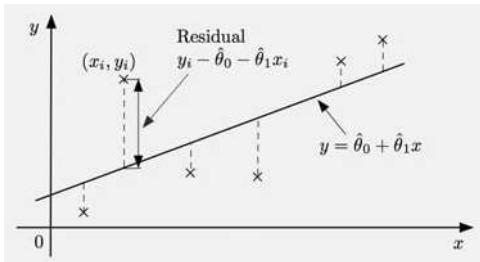
$$\max_{\theta} p_{\Theta|X}(\theta | x) \text{ or } \max_{\theta} \frac{p_X(\theta | x)p_{\Theta}(\theta)}{p_Y(y)}$$
- Sample mean estimate of $\theta = E[X]$

$$\hat{\Theta}_n = (X_1 + \dots + X_n)/n$$
- $1 - \alpha$ confidence interval

$$P(\hat{\Theta}_n^- \leq \theta \leq \hat{\Theta}_n^+) \geq 1 - \alpha, \quad \forall \theta$$
- confidence interval for sample mean
 - let z be s.t. $\Phi(z) = 1 - \alpha/2$

$$P\left(\hat{\Theta}_n - \frac{z\sigma}{\sqrt{n}} \leq \theta \leq \hat{\Theta}_n + \frac{z\sigma}{\sqrt{n}}\right) \approx 1 - \alpha$$

Regression



- Data: $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$
 - Model: $y \approx \theta_0 + \theta_1 x$
- $$\min_{\theta_0, \theta_1} \sum_{i=1}^n (y_i - \theta_0 - \theta_1 x_i)^2 \quad (*)$$
- One interpretation:

$$Y_i = \theta_0 + \theta_1 x_i + W_i, \quad W_i \sim N(0, \sigma^2), \text{ i.i.d.}$$
 - Likelihood function $f_{X,Y|\theta}(x, y; \theta)$ is:
- $$c \cdot \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \theta_0 - \theta_1 x_i)^2 \right\}$$
- Take logs, same as (*)
 - Least sq. \leftrightarrow pretend W_i i.i.d. normal

Linear regression

- Model $y \approx \theta_0 + \theta_1 x$

$$\min_{\theta_0, \theta_1} \sum_{i=1}^n (y_i - \theta_0 - \theta_1 x_i)^2$$
- Solution (set derivatives to zero):

$$\bar{x} = \frac{x_1 + \dots + x_n}{n}, \quad \bar{y} = \frac{y_1 + \dots + y_n}{n}$$

$$\hat{\theta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x}$$
- Interpretation of the form of the solution
 - Assume a model $Y = \theta_0 + \theta_1 X + W$
 W independent of X , with zero mean
 - Check that

$$\hat{\theta}_1 = \frac{\text{cov}(X, Y)}{\text{var}(X)} = \frac{E[(X - E[X])(Y - E[Y])]}{E[(X - E[X])^2]}$$
 - Solution formula for $\hat{\theta}_1$ uses natural estimates of the variance and covariance

The world of linear regression

- **Multiple linear regression:**

- **data:** (x_i, x'_i, x''_i, y_i) , $i = 1, \dots, n$
- **model:** $y \approx \theta_0 + \theta_1 x + \theta'_1 x' + \theta''_1 x''$
- **formulation:**

$$\min_{\theta, \theta', \theta''} \sum_{i=1}^n (y_i - \theta_0 - \theta_1 x_i - \theta'_1 x'_i - \theta''_1 x''_i)^2$$

- **Choosing the right variables**

- model $y \approx \theta_0 + \theta_1 h(x)$
e.g., $y \approx \theta_0 + \theta_1 x^2$
- work with data points $(y_i, h(x))$
- formulation:

$$\min_{\theta} \sum_{i=1}^n (y_i - \theta_0 - \theta_1 h(x_i))^2$$

The world of regression (ctd.)

- **In practice,** one also reports

- Confidence intervals for the θ_i
- “Standard error” (estimate of σ)
- R^2 , a measure of “explanatory power”

- **Some common concerns**

- Heteroskedasticity
- Multicollinearity
- Sometimes misused to conclude causal relations
- etc.

Binary hypothesis testing

- Binary θ ; new terminology:
 - **null hypothesis** H_0 :
 $X \sim p_X(x; H_0)$ [or $f_X(x; H_0)$]
 - **alternative hypothesis** H_1 :
 $X \sim p_X(x; H_1)$ [or $f_X(x; H_1)$]
- Partition the space of possible data vectors
Rejection region R :
 reject H_0 iff data $\in R$
- Types of errors:
 - **Type I (false rejection, false alarm):**
 H_0 true, but rejected
 $\alpha(R) = P(X \in R; H_0)$
 - **Type II (false acceptance, missed detection):**
 H_0 false, but accepted
 $\beta(R) = P(X \notin R; H_1)$

Likelihood ratio test (LRT)

- Bayesian case (MAP rule): choose H_1 if:
 $P(H_1 | X = x) > P(H_0 | X = x)$
 or

$$\frac{P(X = x | H_1)P(H_1)}{P(X = x)} > \frac{P(X = x | H_0)P(H_0)}{P(X = x)}$$
 or

$$\frac{P(X = x | H_1)}{P(X = x | H_0)} > \frac{P(H_0)}{P(H_1)}$$
 (likelihood ratio test)
- Nonbayesian version: choose H_1 if

$$\frac{P(X = x; H_1)}{P(X = x; H_0)} > \xi$$
 (discrete case)

$$\frac{f_X(x; H_1)}{f_X(x; H_0)} > \xi$$
 (continuous case)
- threshold ξ trades off the two types of error
 - choose ξ so that $P(\text{reject } H_0; H_0) = \alpha$
(e.g., $\alpha = 0.05$)

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013 Transcript – Lecture 25

The following content is provided under a Creative Commons license. Your support will help MIT OpenCourseWare continue to offer high quality educational resources for free. To make a donation or view additional materials from hundreds of MIT courses, visit MIT OpenCourseWare at ocw.mit.edu

PROFESSOR: OK, if you have not yet done it, please take a moment to go through the course evaluation website and enter your comments for the class. So what we're going to do today to wrap things up is we're going to go through a tour of the world of hypothesis testing. See a few examples of hypothesis tests, starting from simple ones such as the one the setting that we discussed last time in which you just have two hypotheses, you're trying to choose between them.

But also look at more complicated situations in which you have one basic hypothesis. Let's say that you have a fair coin and you want to test it against the hypotheses that your coin is not fair, but that alternative hypothesis is really lots of different hypothesis. So is my coin fair? Is my die fair? Do I have the correct distribution for random variable, and so on. And I'm going to end up with a few general comments about this whole business.

So the sad thing in simple hypothesis testing problems is the following-- we have two possible models, and this is the classical world so we do not have any prior probabilities on the two hypotheses. Usually we want to think of these hypotheses as not being completely symmetrical, but rather one is the default hypothesis, and usually it's referred to as the null hypothesis. And you want to check whether the null hypothesis is true, whether things are normal as you would have expected them to be, or whether it turns out to be false, in which case an alternative hypothesis would be correct.

So how does one go about it? No matter what approach you use, in the end you're going to end up doing the following. You have the space of all simple observations that you may obtain. So when you do the experiment you're going to get an X vector, a vector of data that's somewhere.

And for some vectors you're going to decide that you accept H_0 . Note for some vectors that you reject H_0 and you accept H_1 . So what you will end up doing is that you're going to have some division of the space of all X 's into two parts, and one part is the rejection region, and one part is the acceptance region. So if you fall in here you accept H_0 , if you fall here you'd reject H_0 .

So to design a hypothesis test basically you need to come up with a division of your X space into two pieces. So the figuring out how to do this involves two elements. One element is to decide what kind of shape so I want for my dividing curve? And having chosen the shape of the dividing curve, where exactly do I put it?

So if you were to cut this space using, let's say, a straight cut you might put it here, or you might put it there, or you might put it there. Where exactly are you going to put it? So let's look at those

two steps. The first issue is to decide the general shape of your rejection region, which is the structure of your test. And the way this is done for the case of two hypothesis is by writing down the likelihood ratio between the two hypothesis.

So let's call that quantity l of X . It's something that you can compute given the data that you have. A high value of l of X basically means that this probability here tends to be bigger than this probability. It means that the data that you have seen are quite likely to have occurred under H_1 , but less likely to have occurred under H_0 .

So if you see data that they are more plausible, can be better explained, under H_1 , then this ratio is big, and you're going to choose in favor of H_1 or reject H_0 . That's what you do if you have discrete data. You use the PMFs. If you have densities, in the case of continues data, again you consider the ratio of the two densities.

So a big l of X is evidence that your data are more compatible with H_1 rather than H_0 . Once you accept this kind of structure then your decision is really made in terms of that single number. That is, you had your data that was some kind of vector, and you condense your data into a single number-- a statistic as it's called-- in this case the likelihood ratio, and you put the dividing point somewhere here call it X_i . And in this region you accept H_1 , in this region you accept H_0 .

So by committing ourselves to using the likelihood ratio in order to carry out the test we have gone from this complicated picture of finding a dividing line in x -space, to a simpler problem of just finding a dividing point on the real line. OK, how are we going?

So what's left to do is to choose this threshold, X_i . Or as it's called, the critical value, for making our decision. And you can place it anywhere, but one way of deciding where to place it is the following-- look at the distribution of this random variable, l of X . It's has a certain distribution under H_0 , and it has some other distribution under H_1 .

If I put my threshold here, here's what's going to happen. When H_0 is true, there is this much probability that I'm going to end up making an incorrect decision. If H_0 is true there's still a probability that my likelihood ratio will be bigger than X_i , and that's the probability of making an incorrect decision of this particular type. That is of making a false rejection of H_0 .

Usually one sets this probability to a certain number, alpha. For example alpha being 5 %. And once you decide that you want this to be 5 %, that determines where this number $\Psi(X_i)$ is going to be.

So the idea here is that I'm going to reject H_0 if the data that I have seen are quite incompatible with H_0 . if they're quite unlikely to have occurred under H_0 . And I take this level, 5%. So I see my data and then I say well if H_0 was true, the probability that I would have seen data of this kind would be less than 5 %.

Given that I saw those data, that suggests that H_0 is not true, and I end up rejecting H_0 . Now of course there's the other type of error probability. If I put my threshold here, if H_1 is true but my

likelihood ratio falls here I'm going to make a mistake of the opposite kind. H_1 is true, but my likelihood ratio turned out to be small, and I decided in favor of H_0 .

This is an error of the other kind, this probability of error we call beta. And you can see that there's a trade-off between alpha and beta. If you move your threshold this way alpha becomes smaller, but beta becomes larger. And the general picture is, in your trade-off, depending on where you put your threshold is as follows-- you can make this beta to be 0 if you put your threshold out here, but in that case you are certain that you're going to make a mistake of the opposite kind. So beta equals 0, alpha equals 1 is one possibility. Beta equals 1 alpha equals 0 is the other possibility if you send your thresholds complete to the other side. And in general you're going to get a trade-off curve of some sort.

And if you want to use a specific value of alpha, for example alpha being 0.05, then that's going to determine for you the probability for beta. Now there's a general, and quite important theorem in statistics, which we're not proving. And which tells us that when we use likelihood ratio tests we get the best possible trade-off curve.

You could think of other ways of making your decisions. Other ways of cutting off your x -space into a rejection and acceptance region. But any other way that you do it is going to end up with some probabilities of error that are going to be above this particular curve.

So the likelihood ratio test turns out to give you the best possible way of dealing with this trade-off between alpha and beta. We cannot minimize alpha and beta simultaneously, there's a trade-off between them. But at least we would like to have a test that deals with this trade-off in the best possible way.

For a given value of alpha we want to have the smallest possible value of beta. And as the theorem is that the likelihood ratio tests do have this optimality property. For a given value of alpha they minimize the probability of error of a different kind.

So let's make all these concrete and look at the simple example. We have two normal distributions with different means. So under H_0 you have a mean of 0. Under H_1 you have a mean of 1. You get your data, you actually get several data drawn from one of the two distributions. And you want to make a decision, which one of the two is true?

So what you do is you write down the likelihood ratio. The density for a vector of data, if that vector was generated according to H_0 -- which is this one, and the density if it was generated according to H_1 . Since we have multiple data the density of a vector is the product of the densities of the individual elements.

Since we're dealing with normals we have those exponential factors. A product of exponentials gives us an exponential of the sum. I'll spare you the details, but this is the form of the likelihood ratio.

The likelihood ratio test tells us that we should calculate this quantity after we get your data, and compare with a threshold. Now you can do some algebra here, and simplify. And by tracing

down the inequalities you're taking logarithms of both sides, and so on. One comes to the conclusion that using a test that has a threshold on this ratio is equivalent to calculating this quantity, and comparing it with a threshold.

Basically this quantity here is monotonic in that quantity. This being larger than the threshold is equivalent to this being larger than the threshold. So this tells us the general structure of the likelihood ratio test in this particular case.

And it's nice because it tells us that we can make our decisions by looking at this simple summary of the data. This quantity, this summary of the data on the basis of which we make our decision is called a statistic. So you take your data, which is a multi-dimensional vector, and you condense it to a single number, and then you make a decision on the basis of that number.

So this is the structure of the test. If I get a large sum of X_i 's this is evidence in favor of H_1 because here the mean is larger. And so I'm going to decide in favor of H_1 or reject H_0 if the sum is bigger than the threshold. How do I choose my threshold? Well I would like to choose my threshold so that the probability of an incorrect decision when H_0 is true the probability of a false rejection equals to a certain number. Alpha, such as for example 5 %.

So you're given here that this is 5 %. You know the distribution of this random variable, it's normal. And you want to find the threshold value that makes this to be true. So this is a type of problem that you have seen several times. You go to the normal tables, and you figure it out. So the sum of the X_i 's has some distribution, it's normal. So that's the distribution of the sum of the X_i 's.

And you want this probability here to be alpha. For this to happen what is the threshold value that makes this to be true? So you know how to solve problems of this kind using the normal tables. A slightly different example is one in which you have two normal distributions that have the same mean -- let's take it to be 0 -- but they have a different variance.

So it's sort of natural that here, if your X 's that you see are kind of big on either side you would choose H_1 . If your X 's are near 0 then that's evidence for the smaller variance you would choose H_0 . So to proceed formally you again write down the form of the likelihood ratio.

So again the density of an X vector under H_0 is this one. It's the product of the densities of each one of the X_i 's. Product of normal densities gives you a product of exponentials, which is exponential of the sum, and that's the expression that you get.

Under the other hypothesis the only thing that changes is the variance. And the variance, in the normal distribution, shows up here in the denominator of the exponent. So you put it there. So this is the general structure of the likelihood ratio test. And now you do some algebra. These terms are constants comparing this ratio to a constant is the same as just comparing the ratio of the exponentials to a constant.

Then you take logarithms, you want to compare the logarithm of this thing to a constant. You do a little bit of algebra, and in the end you find that the structure of the test is to reject H_0 if the sum of the squares of the X_i 's is bigger than the threshold.

So by committing to a likelihood ratio test you are told that you should be making it your decision according to a rule of this type. So this fixes the shape or the structure of the decision region, of the rejection region. And the only thing that's left, once more, is to pick this threshold in order to have the property that the probability of a false rejection is equal to say 5 %.

So that's the probability that H_0 is true, but the sum of the squares accidentally happens to be bigger than my threshold. In which case I end up deciding H_1 . How do I find the value of X_i prime? Well what I need to do is to look at the picture, more or less of this kind, but now I need to look at the distribution of the sum of the X_i 's squared.

Actually the sum of the X_i 's squared is a non-negative random variable. So it's going to have a distribution that's something like this. I look at that distribution, and once more I want this tail probability to be alpha, and that determines where my threshold is going to be. So that's again a simple exercise provided that you know the distribution of this quantity. Do you know it? Well we don't really know it, we have not dealt with this particular distribution in this class. But in principle you should be able to find what it is.

It's a derived distribution problem. You know the distribution of X_i , it's normal. Therefore, by solving a derived distribution problem you can find the distribution of X_i squared. And the X_i squared's are independent of each other, because the X_i 's are independent. So you want to find the distribution of the sum of random variables with known distributions. And since they're independent, in principle, you can do this using the convolution formula.

So in principle, and if you're patient enough, you will be able to find the distribution of this random variable. And then you plot it or tabulate it, and find where exactly is the 95th percentile of that distribution, and that determines your threshold. So this distribution actually turns out to have a nice and simple closed-form formula.

Because this is a pretty common test, people have tabulated that distribution. It's called the chi-square distribution. There's tables available for it. And you look up in the tables, you find the 95th percentile of the distribution, and this way you determine your threshold.

So what's the moral of the story? The structure of the likelihood ratio test tells you what kind of decision region you're going to have. It tells you that for this particular test you should be using the sum of the X_i squared's as your statistic, as the basis for making your decision. And then you need to solve a derived distribution problem to find the probability distribution of your statistic. Find the distribution of this quantity under H_0 , and finally, based on that distribution, after you have derived it, then determine your threshold.

So now let's move on to a somewhat more complicated situation. You have a coin, and you are told that I tried to make a fair coin. Is it fair?

So you have the hypothesis, which is the default-- the null hypothesis-- that the coin is fair. But maybe it isn't. So you have the alternative hypothesis that your coin is not fair. Now what's different in this context is that your alternative hypothesis is not just one specific hypothesis.

Your alternative hypothesis consists of many alternatives. It includes the hypothesis that p is 0.6. It includes the hypothesis that p is 0.51. It includes the hypothesis that p is 0.48, and so on.

So you're testing this hypothesis versus all this family of alternative hypothesis. What you will end up doing is essentially the following-- you get some data. That is, you flip the coin a number of times. Let's say you flip it 1,000 times. You observe some outcome. Let's say you saw 472 heads.

And you ask the question if this hypothesis is true is this value really possible under that hypothesis? Or would it be very much of an outlier? If it looks like an extreme outlier under this hypothesis then I reject it, and I accept the alternative. If this number turns out to be something within the range that you would have expected then you keep, or accept your null hypothesis.

OK so what does it mean to be an outlier or not? First you take your data, and you condense them to a single number. So your detailed data actually would have been a sequence of heads/tails, heads/tails and all that. Any reasonable person would tell you that you shouldn't really care about the exact sequence of heads and tails. Let's just base our decision on the number of heads that we have observed.

So using some kind of reasoning which could be mathematical, or intuitive, or involving artistry-- you pick a one-dimensional, or scalar summary of the data that you have seen. In this case, the summary of the data is just the number of heads that's a quite reasonable one. And so you commit yourself to make a decision on the basis of this quantity.

And you ask the quantity that I'm seeing does it look like an outlier? Or does it look more or less OK? OK, what does it mean to be an outlier? You want to choose the shape of this rejection region, but on the basis of that single number s . And again, the reasonable thing to do in this context would be to argue as follows-- if my coin is fair I expect to see n over 2 heads. That's the expected value.

If the number of heads I see is far from the expected number of heads then I consider this to be an outlier. So if this number is bigger than some threshold X_i . I consider it to be an outlier, and then I'm going to reject my hypothesis.

So we picked our statistic. We picked the general form of how we're going to make our decision, and then we pick a certain significance, or confidence level that we want. Again, this famous 5% number. And we're going to declare something to be an outlier if it lies in the region that has 5% or less probability of occurring.

That is I'm picking my rejection region so that if H_0 is true under the default, or null hypothesis, there's only 5% chance that by accident I fall there, and the thing makes me think that H_1 is going to be true.

So now what's left to do is to pick the value of this threshold. This is a calculation of the usual kind. I want to pick my threshold, my X_i number so that the probability that s is further from the mean by an amount of X_i is less than 5%. Or that the probability of being inside the acceptance region-- so that the distance from the default is less than my threshold. I want that to be 95%.

So this is an equality that you can get using the central limit theorem and the normal tables. There's 95% probability that the number of heads is going to be within 31 from the correct mean. So the way the exercise is done of course, is that we start with this number, 5%. Which translates to this number 95%. And once we have fixed that number then you ask the question what number should we have here to make this equality to be true?

It's again a problem of this kind. You have a quantity whose distribution you know. Why do you know it? The number of heads by the central limit theorem is approximately normal. So this here talks about the normal distribution. You set your alpha to be 5%, and you ask where should I put my threshold so that this probability of being out there is only 5%?

Now in our particular example the threshold turned out to be 31. This number turned out was just 28 away from the correct mean. So these distance was less than the threshold. So we end up not rejecting H_0 .

So we have our rejection region. The way we designed it is that when H_0 is true there's only a small chance, 5%, that we get to data out of there. Data that we would call an outlier. If we see such an outlier we reject H_0 . If what we see is not an outlier as in this case, where that distance turned out to be kind of small, then we do not reject H_0 .

An interesting little piece of language here, people generally prefer to use this terminology-- to say that H_0 is not rejected by the data. Instead of saying that H_0 is accepted. In some sense they're both saying the same thing, but the difference is sort of subtle. When I say not rejected what I mean is that I got some data that are compatible with my hypothesis.

That is the data that I got do not falsify the hypothesis that I had, my null hypothesis. So my null hypothesis is still alive, and may be true. But from data you can never really prove that the hypothesis is correct. Perhaps my coin is not fair in some other complicated way.

Perhaps I was just lucky, and even though my coin is not fair I ended up with an outcome that suggests that it's fair. Perhaps my coin flips are not independent as I assumed in my model. So there's many ways that my null hypothesis could be wrong, and still I got data that tells me that my hypothesis is OK.

So this is the general way that things work in science. One comes up with a model or a theory. This is the default theory, and we work with that theory trying to find whether there are examples that violate the theory. If you find data and examples that violate the theory your theory is falsified, and you need to look for a new one.

But when you have your theory, really no amount of data can prove that your theory is correct. So we have the default theory that the speed of light is constant as long as we do not find any

data that runs counter to it. We stay with that theory, but there's no way of really proving this, no matter how many experiments we do.

But there could be experiments that falsify that theory, in which case we need to do look for a new one. So there's a bit of an asymmetry here in how we treat the alternative hypothesis. H_0 is the default which we'll accept until we see some evidence to the contrary. And if we see some evidence to the contrary we reject it. As long as we do not see evidence to the contrary then we keep working with it, but always take it with a grain of salt.

You can never really prove that a coin has a bias exactly equal to $1/2$. Maybe the bias is equal to 0.50001 , so the bias is not $1/2$. But with an experiment with 1,000 coin tosses you wouldn't be able to see this effect.

OK, so that's how you go about testing about whether your coin is fair. You can also think about testing whether a die is fair. So for a die the null hypothesis would be that every possible result when you roll the die has equal probability and equal to $1/6$. And you also make the hypothesis that your die rolls are statistically independent from each other.

So I take my die, I roll it a number of times, little n , and I count how many 1's I got, how many 2's I got, how many 3's I got, and these are my data. I count how many times I observed a specific result in my die roll that was equal to sum i .

And now I ask the question-- the N_i 's that I observed, are they compatible with my hypothesis or not? What does compatible to my hypothesis mean? Under the null hypothesis N_i should be approximately equal, or is equal in expectation to N times little P_i . And in our example this little P_i is of course $1/6$.

So if my die is fair the number of ones I expect to see is equal to the number of rolls times $1/6$. The number of 2's I expect to see is again that same number. Of course there's randomness, so I do not expect to get exactly that number. But I can ask how far away from the expected values was i ?

If my capital N_i 's turn to be very different from $N/6$ this is evidence that my die is not fair. If those numbers turn out to be close to N times $1/6$ then I'm going to say there's no evidence that would lead me to reject this hypothesis. So this hypothesis remains alive.

So someone has come up with this thought that maybe the right statistic to use, or the right way of quantifying how far away are the N_i 's from their mean is to look at this quantity. So I'm looking at the expected value of N_i under the null hypothesis. See what I got, take the square of this, and add it over all i 's.

But also throw in these terms in the denominator. And why that term is there, that's a longer story. One can write down certain likelihood ratios, do certain Taylor Series approximations, and there's a Heuristic argument that justifies why this would be a good form for the test to use.

So there's a certain art that's involved in this step that some people somehow decided that it's a reasonable thing to do is to calculate. Once you get your results to calculate this one-dimensional summary of your result, this is going to be your statistic, and compare that statistic to a threshold. And that's how you make your decision.

So by this point we have fixed the type of the rejection region that we're going to have. So we've chosen the qualitative structure of our test, and the only thing that's now left is to choose the particular threshold we're going to use. And the recipe, once more, is the same.

We want to set our threshold so that the probability of a false rejection is 5%. We want the probability that our data fall in here is only 5% when the null hypothesis is true. So that's the same as setting our threshold X_i so that the probability that our test statistic is bigger than that threshold. We want that probability to be only 0.05.

So to solve a problem of this kind what is it that you need to do? You need to find the probability distribution of capital T. So once more it's the same picture. You need to do some calculations of some sort, and come up with the distribution of the random variable T, where T is defined this way. You want to find this distribution under hypothesis H_0 .

Once you find what that distribution is then you can solve this usual problem. I want this probability here to be 5%. What should my threshold be? So what does this boil down to? Finding the distribution of capital T is in some sense a messy, difficult, derived distribution problem. From this model we know the distribution of the capital N_i 's. And actually we can even write down the joint distribution of the capital N_i 's.

In fact we can make an approximation here. Capital N_i is a binomial random variable. Let's say the number of 1's that I got in little N rolls off my die. So that's a binomial random variable. When little n is big this is going to be approximately normal. So we have normal random variables, or approximately normal minus a constant. They're still approximately normal. We take the squares of these, scale them so you can solve a derived distribution problem to find the distribution of this quantity.

You can do more work, more derived distribution work, and find the distribution of capital T. So this is a tedious matter, but because this test is used quite often, again people have done those calculations. They have found the distribution of capital T, and it's available in tables. And you go to those tables, and you find the appropriate threshold for making a decision of this type.

Now to give you a sense of how complicated hypothesis one might have to deal with let's make things one level more complicated. So here you can think this X is a discrete random variable. This is the outcome of my roll. And I had a model in which the possible values of my discrete random variables they have probabilities all equal to 1/6.

So my null hypothesis here was a particular PMF for the random variable capital X. So another way of phrasing what happened in this problem was the question is my PMF correct? So this is the PMF of the result of one die roll. You're asking the question is my PMF correct? Make it more complicated.

How about the question of the type is my PDF correct when I have continuous data? So I have hypothesized that's the probability distribution that I have is let's say a particular normal. I get lots of results from that random variable. Can I tell whether my results look like normal or not? What are some ways of going about it?

Well, we saw in the previous slide that there is a methodology for deciding if your PMF is correct. So you could take your normal results, the data that you got from your experiment, and discretize them, and so now you're dealing with discrete data. And sort of used in previous methodology to solve a discrete problem of the type is my PDF correct?

So in practice the way this is done is that you get all your data, let's say data points of this kind. You split your space into bins, and you count how many you have in each bin. So you get this, and that, and that, and nothing. So that's a histogram that you get from the data that you have. Like the very familiar histograms that you see after each one of our quizzes.

So if you look at these histogram, and you ask does it look like normal? OK, we need a systematic way of going about it. If it were normal you can calculate the probability of falling in this interval. The probability of falling in that interval, probability of falling into that interval. So you would have expected values of how many results, or data points, you would have in this interval. And compare these expected values for each interval with the actual ones that you observed. And then take the sum of squares, and so on, exactly as in the previous slide. And this gives you a way of going about it.

This is a little messy. It gets hard to do because you have the difficult decision of how do you choose the bin size? If you take your bins to be very narrow you would get lots of bins with 0's, and a few bins that only have one outcome in them. It probably wouldn't feel right. If you choose your bins to be very wide then you're losing a lot of information. Is there some way of making a test without creating bins?

This is just to illustrate the clever ideas of what statisticians have thought about. And here's a really cute way of going about a test, whether my distribution is correct or not. Here we're essentially plotting a PMF, or an approximation of a PDF. And we ask does it look like the PDF we assumed?

Instead of working with PDFs let's work with cumulative distribution functions. So how does this go? The true normal distribution that I have hypothesized, the density that I'm hypothesizing-- my null hypothesis-- has a certain CDF that I can plot. So supposed that my hypothesis H_0 is that the X's are normal with our standard normals, and I plot the CDF of the standard normal, which is the sort of continuous looking curve here.

Now I get my data, and I plot the empirical CDF. What's the empirical CDF? In the empirical CDF you ask the question what fraction of the data fell below 0? You get a number. What fraction of my data fell below 1? I get a number. What fraction of my data fell below 2, and so on.

So you're talking about fractions of the data that fell below each particular number. And by plotting those fractions as a function of this number you get something that looks like a CDF. And it's the CDF suggested by the data.

Now the fraction of the data that fall below 0 in my experiment is-- if my hypothesis were true-- expected to be 1/2. 1/2 is the value of the true CDF. I look at the fraction that I got, it's expected to be that number. But there's randomness, so it's might be a little different than that. For any particular value, the fraction that I got below a certain number-- the fraction of data that we're below, 2, its expectation is the probability of falling below 2, which is the correct CDF.

So if my hypothesis is true the empirical CDF that I get based on data should, when n is large, be very close to the true CDF. So a way of judging whether my model is correct or not is to look at the assumed CDF, the CDF under hypothesis H_0 . Look at the CDF that I constructed based on the data, and see whether they're close enough or not.

And by close enough, I mean I'm going to look at all the possible X 's, and look at the maximum distance between those two curves. And I'm going to have a test that decides in favor of H_0 if this distance is small, and in favor of H_1 if this distance is large.

That still leaves me the problem of coming up with a threshold. Where exactly do I put my threshold? Because this test is important enough, and is used frequently people have made the effort to try to understand the probability distribution of this quite difficult random variable. One needs to do lots of approximations and clever calculations, but these have led to values and tabulated values for the probability distribution of this random variable.

And, for example, those tabulated values tell us that if we want 5% false rejection probability, then our threshold should be 1.36 divided by the square root of n . So we know where to put our threshold for this particular value. If we want this particular error or error probability to occur.

So that's about as hard and sophisticated classical statistics get. You want to have tests for hypotheses that are not so easy to handle. People somehow think of clever ways of doing tests of this kind. How to compare the theoretical predictions with the observed predictions with the observed data. Come up with some measure of the difference between theory and data, and if that difference is big, than you reject your hypothesis.

OK, of course that's not the end of the field of statistics, there's a lot more. In some ways, as we kept moving through today's lecture, the way that we constructed those rejection regions was more and more ad hoc. I pulled out of a hat a particular measure of fit between data and the model. And I said let's just use a test based on this.

There are attempts at more or less systematic ways of coming up with the general shape of rejection regions that have at least some desirable or favorable theoretical properties. Some more specific problems that people study-- instead of having a test, is this the correct PDF? Yes or no. I just give you data, and I ask you tell me, give me a model or a PDF for those data.

OK, my thoughts of this kind are of many types. One general method is you form a histogram, and then you take your histogram and plot a smooth line, that kind of fits the histogram. This still leaves the question of how do you choose the bins? The bin size in your histograms. How narrow do you take them? And that depends on how many data you have, and there's a lot of theory that tells you about the best way of choosing the bin sizes, and the best ways of smoothing the data that you have.

A completely different topic is in signal processing -- you want to do your inference. Not only you want it to be good, but you also want it to be fast in a computational way. You get data in real time, lots of data. You want to keep processing and revising your estimates and your decisions as they come and go.

Another topic that was briefly touched upon the last couple of lectures is that when you set up a model, like a linear regression model, you choose some explanatory variables, and you try to predict y from your X , these variables. You have a choice of what to take as your explanatory variables. Are there systematic ways of picking the right X variables to try to estimate a Y .

For example should I try to estimate Y on the basis of X ? Or on the basis of X -squared? How do I decide between the two?

Finally, the rage these days has to do with anything big, high-dimensional. Complicated models of complicated things, and tons and tons of data. So these days data are generated everywhere. The amounts of data are humongous. Also, the problems that people are interested in tend to be very complicated with lots of parameters.

So I need specially tailored methods that can give you good results, or decent results even in the face of these huge amounts of data, and possibly with computational constraints. So with huge amounts of data you want methods that are simple, but still can deliver for you meaningful answers.

Now as I mentioned some time ago, this whole field of statistics is very different from the field of probability. In some sense all that we're doing in statistics is probabilistic calculations. That's what the theory kind of does. But there's a big element of art.

You saw that we chose the shape of some decision regions or rejection regions in a somewhat ad hoc way. There's even more basic things. How do you organize your data? How do you think about which hypotheses you would like to test, and so on. There's a lot of art that's involved here, and there's a lot that can go wrong.

So I'm going to close with a note that you can take either as pessimistic or optimistic. There is a famous paper that came out a few years ago and has been cited about a 1,000 times or so. And the title of the paper is Why Most Published Research Findings Are False. And it's actually a very good argument why, in fields like psychology or the medical science and all that a lot of what you see published-- that yes, this drug has an effect on that particular disease-- is actually false, because people do not do their statistics correctly.

There's lots of biases in what people do. I mean an obvious bias is that you only published a result when you see something. So the null hypothesis is that the drug doesn't work. You do your tests, the drug didn't work, OK, you just go home and cry.

But if by accident that 5% happens, and even though the drug doesn't work, you got some outlier data, and it seemed to be working. Then you're excited, you publish it. So that's clearly a bias. That gets results to be published, even though they do not have a solid foundation behind them.

Then there's another thing, OK? I'm picking my 5%. So H_0 is true there's a small probability that the data will look like an outlier, and in that case I published my result. OK it's only 5% -- it's not going to happen too often. But suppose that I go and do a 1,000 different tests? Test H_0 against this hypothesis, test H_0 against that hypothesis , test H_0 against that hypothesis.

Some of these tests, just by accident might turn out to be in favor of H_1 , and again these are selected to be published. So if you do lots and lots of tests and in each one you have a 5% probability of error, when you consider the collection of all those tests, actually the probability of making incorrect inferences is a lot more than 5%.

One basic principle in being systematic about such studies is that you should first pick your hypothesis that you're going to test, then get your data, and do your hypothesis testing. What would be wrong is to get your data, look at them, and say OK I'm going now to test for these 100 different hypotheses, and I'm going to choose my hypothesis to be for features that look abnormal in my data.

Well, given enough data, you can always find some abnormalities just by chance. And if you choose to make a statistical test-- is this abnormality present? Yes, it will be present. Because you first found the abnormality, and then you tested for it. So that's another way that things can go wrong.

So the moral of this story is that while the world of probability is really beautiful and solid, you have your axioms. Every question has a unique answer that by now you can, all of you, find in a very reliable way. Statistics is a dirty and difficult business. And that's why the subject is not over. And if you're interested in it, it's worth taking follow-on courses in that direction. OK so have good luck in the final, do well, and have a nice vacation afterwards.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Problem Set 11
Never Due
Covered on Final Exam

1. Problem 7, page 509 in textbook

Derive the ML estimator of the parameter of a Poisson random variable based of i.i.d. observations X_1, \dots, X_n . Is the estimator unbiased and consistent?

2. Caleb builds a particle detector and uses it to measure radiation from far stars. On any given day, the number of particles Y that hit the detector is conditionally distributed according to a Poisson distribution conditioned on parameter x . The parameter x is unknown and is modeled as the value of a random variable X , exponentially distributed with parameter μ as follows.

$$f_X(x) = \begin{cases} \mu e^{-\mu x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Then, the conditional PDF of the number of particles hitting the detector is,

$$p_{Y|X}(y | x) = \begin{cases} \frac{e^{-x} x^y}{y!} & y = 0, 1, 2, \dots \\ 0 & \text{otherwise} \end{cases}$$

- (a) Find the MAP estimate of X from the observed particle count y .
 (b) Our goal is to find the conditional expectation estimator for X from the observed particle count y .

i. Show that the posterior probability distribution for X given Y is of the form

$$f_{X|Y}(x | y) = \frac{\lambda^{y+1}}{y!} x^y e^{-\lambda x}, \quad x > 0$$

and find the parameter λ . You may find the following equality useful (it is obviously true if the equation above describes a true PDF):

$$\int_0^\infty a^{y+1} x^y e^{-ax} dx = y! \quad \text{for any } a > 0$$

- ii. Find the conditional expectation estimate of X from the observed particle count y .
Hint: you might want to express $x f_{X|Y}(x | y)$ in terms of $f_{X|Y}(x | y + 1)$.
 (c) Compare the two estimators you constructed in part (a) and part (b).
3. Consider a Bernoulli process X_1, X_2, X_3, \dots with unknown probability of success q . Define the k th inter-arrival time T_k as

$$T_1 = Y_1, \quad T_k = Y_k - Y_{k-1}, \quad k = 2, 3, \dots$$

where Y_k is the time of the k th success. This problem explores estimation of q from observed inter-arrival times $\{t_1, t_2, t_3, \dots\}$. In problem set 10, we solved the problem using Bayesian inference. Our focus here will be on classical estimation.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
 (Fall 2010)

We assume that q is an unknown parameter in the interval $(0, 1]$. Denote the true parameter by q^* . Denote by \hat{Q}_k the maximum likelihood estimate (MLE) of q given k recordings, $T_1 = t_1, \dots, T_k = t_k$.

- (a) Compute \hat{Q}_k . Is this different from the MAP estimate you found in problem set 10?
- (b) Show that for all $\epsilon > 0$

$$\lim_{k \rightarrow \infty} \mathbf{P} \left(\left| \frac{1}{\hat{Q}_k} - \frac{1}{q^*} \right| > \epsilon \right) = 0$$

- (c) Assume $q^* \geq 0.5$. Give a lower bound on k such that

$$\mathbf{P} \left(\left| \frac{1}{\hat{Q}_k} - \frac{1}{q^*} \right| \leq 0.1 \right) \geq 0.95$$

4. A body at temperature θ radiates photons at a given wavelength. This problem will have you estimate θ , which is fixed but unknown. The PMF for the number of photons K in a given wavelength range and a fixed time interval of one second is given by,

$$p_K(k; \theta) = \frac{1}{Z(\theta)} e^{-\frac{k}{\theta}}, k = 0, 1, 2, \dots$$

$Z(\theta)$ is a normalization factor for the probability distribution (the physicists call it the partition function). You are given the task of determining the temperature of the body to two significant digits by photon counting in non-overlapping time intervals of duration one second. The photon emissions in non-overlapping time intervals are statistically independent from each other.

- (a) Determine the normalization factor $Z(\theta)$.
- (b) Compute the expected value of the photon number measured in any 1 second time interval, $\mu_K = \mathbf{E}_\theta[K]$, and its variance, $\text{var}_\theta(K) = \sigma_K^2$.
- (c) You count the number k_i of photons detected in n non-overlapping 1 second time intervals. Find the maximum likelihood estimator, $\hat{\Theta}_n$, for temperature Θ . Note, it might be useful to introduce the average photon number $s_n = \frac{1}{n} \sum_{i=1}^n k_i$. In order to keep the analysis simple we assume that the body is hot, i.e. $\theta \gg 1$.
 You may use the approximation: $\frac{1}{e^{\frac{1}{\theta}} - 1} \approx \theta$ for $\theta \gg 1$.

In the following questions we wish to estimate the mean of the photon count in a one second time interval using the estimator \hat{K} , which is given by,

$$\hat{K} = \frac{1}{n} \sum_{i=1}^n K_i.$$

- (d) Find the number of samples n for which the noise to signal ratio for \hat{K} , (i.e., $\frac{\sigma_{\hat{K}}}{\mu_{\hat{K}}}$), is 0.01.
 - (e) Find the 95% confidence interval for the mean photon count estimate for the situation in part (d). (You may use the central limit theorem.)
5. The RandomView window factory produces window panes. After manufacturing, 1000 panes were loaded onto a truck. The weight W_i of the i -th pane (in pounds) on the truck is modeled as a random variable, with the assumption that the W_i 's are independent and identically distributed.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
 (Fall 2010)

- (a) Assume that the measured weight of the load on the truck was 2340 pounds, and that $\text{var}(W_i) \leq 4$. Find an approximate 95 percent confidence interval for $\mu = \mathbf{E}[W_i]$, using the Central Limit Theorem.
- (b) Now assume instead that the random variables W_i are i.i.d., with an exponential distribution with parameter $\theta > 0$, i.e., a distribution with PDF

$$f_W(w; \theta) = \theta e^{-\theta w}.$$

What is the maximum likelihood estimate of θ , given that the truckload has weight 2340 pounds?

6. Given the five data pairs (x_i, y_i) in the table below,

x	0.8	2.5	5	7.3	9.1
y	-2.3	20.9	103.5	215.8	334

we want to construct a model relating x and y . We consider a linear model

$$Y_i = \theta_0 + \theta_1 x_i + W_i, \quad i = 1, \dots, 5,$$

and a quadratic model

$$Y_i = \beta_0 + \beta_1 x_i^2 + V_i, \quad i = 1, \dots, 5.$$

where W_i and V_i represent additive noise terms, modeled by independent normal random variables with mean zero and variance σ_1^2 and σ_2^2 , respectively.

- (a) Find the ML estimates of the linear model parameters.
- (b) Find the ML estimates of the quadratic model parameters.

Note: You may use the regression formulas and the connection with ML described in pages 478-479 of the text. However, the regression material is outside the scope of the final.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.04F Probabilistic Systems Analysis and Applied Probability

Fall 2010

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Problem Set 11 Solutions

1. Check book solutions.
2. (a) To find the MAP estimate, we need to find the value x that maximizes the conditional density $f_{X|Y}(x | y)$ by taking its derivative and setting it to 0.

$$\begin{aligned} f_{X|Y}(x | y) &= \frac{p_{Y|X}(y | x) \cdot f_X(x)}{p_Y(y)} \\ &= \frac{e^{-x} x^y}{y!} \cdot \mu e^{-\mu x} \cdot \frac{1}{p_Y(y)} \\ &= \frac{\mu}{y! p_Y(y)} \cdot e^{-(\mu+1)x} x^y \end{aligned}$$

$$\begin{aligned} \frac{d}{dx} f_{X|Y}(x | y) &= \frac{d}{dx} \left(\frac{\mu}{y! p_Y(y)} \cdot e^{-(\mu+1)x} x^y \right) \\ &= \frac{\mu}{y! p_Y(y)} x^{y-1} e^{-(\mu+1)x} (y - x(\mu + 1)) \end{aligned}$$

Since the only factor that depends on x which can take on the value 0 is $(y - x(\mu + 1))$, the maximum is achieved at

$$\hat{x}_{\text{MAP}}(y) = \frac{y}{1 + \mu}$$

It is easy to check that this value is indeed maximum (the first derivative changes from positive to negative at this value).

- (b) i. To show the given identity, we need to use Bayes' rule. We first compute the denominator, $p_Y(y)$

$$\begin{aligned} p_Y(y) &= \int_0^\infty \frac{e^{-x} x^y}{y!} \mu e^{-\mu x} dx \\ &= \frac{\mu}{y! (1 + \mu)^{y+1}} \int_0^\infty (1 + \mu)^{y+1} x^y e^{-(1+\mu)x} dx \\ &= \frac{\mu}{(1 + \mu)^{y+1}} \end{aligned}$$

Then, we can substitute into the equation we had derived in part (a)

$$\begin{aligned} f_{X|Y}(x | y) &= \frac{\mu}{y! p_Y(y)} x^y e^{-(\mu+1)x} \\ &= \frac{\mu}{y!} \frac{(1 + \mu)^{y+1}}{\mu} x^y e^{-(\mu+1)x} \\ &= \frac{(1 + \mu)^{y+1}}{y!} x^y e^{-(\mu+1)x} \end{aligned}$$

Thus, $\lambda = 1 + \mu$.

ii. We first manipulate $xf_{X|Y}(x | y)$:

$$\begin{aligned} xf_{X|Y}(x | y) &= \frac{(1+\mu)^{y+1}}{y!} x^{y+1} e^{-(\mu+1)x} \\ &= \frac{y+1}{1+\mu} \frac{(1+\mu)^{y+2}}{(y+1)!} x^{y+1} e^{-(\mu+1)x} \\ &= \frac{y+1}{1+\mu} f_{X|Y}(x | y+1) \end{aligned}$$

Now we can find the conditional expectation estimator:

$$\begin{aligned} \hat{x}_{\text{CE}}(y) &= \mathbf{E}[X | Y = y] = \int_0^\infty xf_{X|Y}(x | y) dx \\ &= \int_0^\infty \frac{y+1}{1+\mu} f_{X|Y}(x | y+1) dx = \frac{y+1}{1+\mu} \end{aligned}$$

- (c) The conditional expectation estimator is always higher than the MAP estimator by $\frac{1}{1+\mu}$.
3. (a) The likelihood function is

$$\prod_{i=1}^k P_{T_i}(T_i = t_i | Q = q) = q^k (1-q)^{\sum_i^k t_i - k}.$$

To maximize the above probability we set its derivative with respect to q to zero

$$kq^{k-1}(1-q)^{\sum_i^k t_i - k} - (\sum_i^k t_i - k)q^k(1-q)^{\sum_i^k t_i - k - 1} = 0,$$

or equivalently

$$k(1-q) - (\sum_i^k t_i - k)q = 0,$$

which yields $\hat{Q}_k = \frac{k}{\sum_{i=1}^k t_i}$. This is not different from the MAP estimate found before. Since the MAP estimate is calculated using a uniform prior, the likelihood function is a ‘scaled’ version of posterior probability and they can be maximized at the same value of q .

- (b) Since $\frac{1}{\hat{Q}_k} = \frac{\sum_{i=1}^k T_i}{k}$, and that each T_i is independent identically distributed, it follows that $\frac{1}{\hat{Q}_k}$ is actually a sample mean estimator. The weak law of large numbers says that, when the number of samples increases to infinity, the sample mean estimator converges to the actual mean, which is $\frac{1}{q^*}$ in this case. So we can write the limit of probability as

$$\lim_{k \rightarrow \infty} \mathbf{P} \left(\left| \frac{1}{\hat{Q}_k} - \frac{1}{q^*} \right| > \epsilon \right) = \lim_{k \rightarrow \infty} \mathbf{P} \left(\left| \frac{\sum_{i=1}^k T_i}{k} - \mathbf{E}[T_1] \right| > \epsilon \right) = 0.$$

(c) Chebyshev inequality states that

$$\mathbf{P} \left(\left| \frac{\sum_{i=1}^k T_i}{k} - \mathbf{E}[T_1] \right| \geq \epsilon \right) \leq \frac{\text{var}(T_1)}{k\epsilon^2}.$$

So we have

$$\begin{aligned} \mathbf{P} \left(\left| \frac{1}{\hat{Q}_k} - \frac{1}{q^*} \right| \leq 0.1 \right) &= \mathbf{P} \left(\left| \frac{\sum_{i=1}^k T_i}{k} - \frac{1}{q^*} \right| \leq 0.1 \right) \\ &= 1 - \mathbf{P} \left(\left| \frac{\sum_{i=1}^k T_i}{k} - \mathbf{E}[T_1] \right| \geq 0.1 \right) \geq 1 - \frac{\text{var}(T_1)}{k * 0.1^2} \end{aligned}$$

To ensure the above probability to be greater than 0.95, we need that

$$1 - \frac{\text{var}(T_1)}{k * 0.1^2} = 1 - \frac{\frac{1-q}{q^2}}{k * 0.1^2} \geq 0.95,$$

or

$$k \geq 2000\text{var}(T_1) = 2000 \frac{1-q}{q^2}$$

The number of observations k needed depends on the variance of T_1 . For q close to 1, the variance is close to 0, and the required number of observations is very small (close to 0). For $q = 1/2$, the variance is maximum ($\text{var}(T_1) = 2$), and we require $k = 4000$. Thus, to guarantee the required accuracy and confidence for all q , we need that,

$$k \geq 4000.$$

4. (a) Normalization of the distribution requires:

$$1 = \sum_{k=0}^{\infty} p_K(k; \theta) = \sum_{k=0}^{\infty} \frac{e^{-\frac{k}{\theta}}}{Z(\theta)} = \frac{1}{Z(\theta)} \sum_{k=0}^{\infty} e^{-\frac{k}{\theta}} = \frac{1}{Z(\theta) \cdot (1 - e^{-\frac{1}{\theta}})},$$

$$\text{so } Z(\theta) = \frac{1}{1 - e^{-\frac{1}{\theta}}}.$$

(b) Rewriting $p_K(k; \theta)$ as:

$$p_K(k; \theta) = \left(e^{-\frac{1}{\theta}} \right)^k \left(1 - e^{-\frac{1}{\theta}} \right), \quad k = 0, 1, \dots$$

the probability distribution for the photon number is a geometric probability distribution with probability of success $p = 1 - e^{-\frac{1}{\theta}}$, and it is shifted with 1 to the left since it starts with $k = 0$. Therefore the photon number expectation value is

$$\mu_K = \frac{1}{p} - 1 = \frac{1}{1 - e^{-\frac{1}{\theta}}} - 1 = \frac{1}{e^{\frac{1}{\theta}} - 1}$$

and its variance is

$$\sigma_K^2 = \frac{1-p}{p^2} = \frac{e^{-\frac{1}{\theta}}}{(1 - e^{-\frac{1}{\theta}})^2} = \mu_K^2 + \mu_K.$$

- (c) The joint probability distribution for the k_i is

$$p_K(k_1, \dots, k_n; \theta) = \frac{1}{Z(\theta)^n} \prod_{i=1}^n e^{-k_i/\theta} = \frac{1}{Z(\theta)^n} e^{-\frac{1}{\theta} \sum_{i=1}^n k_i}.$$

The log likelihood is $-n \cdot \log Z(\theta) - 1/\theta \sum_{i=1}^n k_i$.

We find the maxima of the log likelihood by setting the derivative with respect to the parameter θ to zero:

$$\frac{d}{d\theta} \log p_K(k_1, \dots, k_n; \theta) = -n \cdot \frac{e^{-\frac{1}{\theta}}}{\theta^2(1 - e^{-\frac{1}{\theta}})} + \frac{1}{\theta^2} \sum_{i=1}^n k_i = 0$$

or

$$\frac{1}{e^{\frac{1}{\theta}} - 1} = \frac{1}{n} \sum_{i=1}^n k_i = s_n.$$

For a hot body, $\theta \gg 1$ and $\frac{1}{e^{\frac{1}{\theta}} - 1} \approx \theta$, we obtain

$$\theta \approx \frac{1}{n} \sum_{i=1}^n k_i = s_n.$$

Thus the maximum likelihood estimator $\hat{\Theta}_n$ for the temperature is given in this limit by the sample mean of the photon number

$$\hat{\Theta}_n = \frac{1}{n} \sum_{i=1}^n K_i.$$

- (d) According to the central limit theorem, the sample mean approaches for large n a Gaussian distribution with standard deviation our root mean square error

$$\sigma_{\hat{\Theta}_n} = \frac{\sigma_K}{\sqrt{n}}.$$

To allow only for 1% relative root mean square error in the temperature, we need $\frac{\sigma_K}{\sqrt{n}} < 0.01\mu_K$. With $\sigma_K^2 = \mu_K^2 + \mu_K$ it follows that

$$\sqrt{n} > \frac{\sigma_K}{0.01\mu_K} = 100 \frac{\sqrt{\mu_K^2 + \mu_K}}{\mu_K} = 100 \sqrt{1 + \frac{1}{\mu_K}}.$$

In general, for large temperatures, i.e. large mean photon numbers $\mu_K \gg 1$, we need about 10,000 samples.

- (e) The 95% confidence interval for the temperature estimate for the situation in part (d), i.e.

$$\sigma_{\hat{\Theta}_n} = \frac{\sigma_K}{\sqrt{n}} = 0.01\mu_K,$$

is

$$[\hat{K} - 1.96\sigma_{\hat{K}}, \hat{K} + 1.96\sigma_{\hat{K}}] = [\hat{K} - 0.0196\mu_K, \hat{K} + 0.0196\mu_K].$$

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
 (Fall 2010)

5. (a) The sample mean estimator $\hat{\Theta}_n = \frac{W_1 + \dots + W_n}{n}$ in this case is

$$\hat{\Theta}_{1000} = \frac{2340}{1000} = 2.34.$$

From the standard normal table, we have $\Phi(1.96) = 0.975$, so we obtain

$$\mathbf{P}\left(\frac{|\hat{\Theta}_{1000} - \mu|}{\sqrt{\text{var}(W_i)/1000}} \leq 1.96\right) \approx 0.95.$$

Because the variance is less than 4, we have

$$\mathbf{P}\left(\hat{\Theta}_{1000} - \mu \leq 1.96\sqrt{\text{var}(W_i)/1000}\right) \leq \mathbf{P}\left(\hat{\Theta}_{1000} - \mu \leq 1.96\sqrt{4/1000}\right),$$

and letting the right-hand side of the above equation ≈ 0.95 gives a 95% confidence, i.e.,

$$[\hat{\Theta}_{1000} - 1.96\sqrt{4/1000}, \hat{\Theta}_{1000} + 1.96\sqrt{4/1000}] = [\hat{\Theta}_{1000} - 0.124, \hat{\Theta}_{1000} + 0.124] = [2.216, 2.464]$$

- (b) The likelihood function is

$$f_W(w; \theta) = \prod_{i=1}^n f_{W_i}(w_i; \theta) = \prod_{i=1}^n \theta e^{-\theta w_i},$$

And the log-likelihood function is

$$\log f_W(w; \theta) = n \log \theta - \theta \sum_{i=1}^n w_i,$$

The derivative with respect to θ is $\frac{n}{\theta} - \sum_{i=1}^n w_i$, and by setting it to zero, we see that the maximum of $\log f_W(w; \theta)$ over $\theta \geq 0$ is attained at $\hat{\theta}_n = \frac{n}{\sum_{i=1}^n w_i}$. The resulting estimator is

$$\hat{\Theta}_n^{mle} = \frac{n}{\sum_{i=1}^n W_i}.$$

In this case,

$$\hat{\Theta}_n^{mle} = \frac{1000}{2340} = 0.4274.$$

6. (a) Using the regression formulas of Section 9.2, we have

$$\hat{\theta}_1 = \frac{\sum_{i=1}^5 (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^5 (x_i - \bar{x})^2}, \quad \hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x},$$

where

$$\bar{x} = \frac{1}{5} \sum_{i=1}^5 x_i = 4.94, \quad \bar{y} = \frac{1}{5} \sum_{i=1}^5 y_i = 134.38.$$

The resulting ML estimates are

$$\hat{\theta}_1 = 40.53, \quad \hat{\theta}_0 = -65.86.$$

(b) Using the same procedure as in part (a), we obtain

$$\hat{\theta}_1 = \frac{\sum_{i=1}^5 (x_i^2 - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^5 (x_i^2 - \bar{x})^2}, \quad \hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x},$$

where

$$\bar{x} = \frac{1}{5} \sum_{i=1}^5 x_i^2 = 33.60, \quad \bar{y} = \frac{1}{5} \sum_{i=1}^5 y_i = 134.38.$$

which for the given data yields

$$\hat{\theta}_1 = 4.09, \quad \hat{\theta}_0 = -3.07.$$

Figure 1 shows the data points (x_i, y_i) , $i = 1, \dots, 5$, the estimated linear model

$$y = 40.53x - 65.86,$$

and the estimated quadratic model

$$y = 4.09x^2 - 3.07.$$

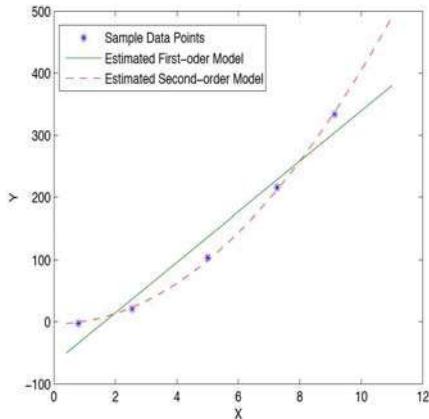


Figure 1: Regression Plot

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

LECTURE 25

Outline

- **Reference:** Section 9.4

- Review of simple binary hypothesis tests
 - examples
- Testing composite hypotheses
 - is my coin fair?
 - is my die fair?
 - goodness of fit tests

Simple binary hypothesis testing

- **null hypothesis** H_0 :
 $X \sim p_X(x; H_0)$ [or $f_X(x; H_0)$]
- **alternative hypothesis** H_1 :
 $X \sim p_X(x; H_1)$ [or $f_X(x; H_1)$]
- Choose a **rejection region** R ;
 reject H_0 iff data $\in R$
- Likelihood ratio test: reject H_0 if

$$\frac{p_X(x; H_1)}{p_X(x; H_0)} > \xi \quad \text{or} \quad \frac{f_X(x; H_1)}{f_X(x; H_0)} > \xi$$
- fix false rejection probability α
 (e.g., $\alpha = 0.05$)
- choose ξ so that $P(\text{reject } H_0; H_0) = \alpha$

Example (test on normal mean)

- n data points, i.i.d.
 $H_0: X_i \sim N(0, 1)$
 $H_1: X_i \sim N(1, 1)$
- Likelihood ratio test; rejection region:

$$\frac{(1/\sqrt{2\pi})^n \exp\{-\sum_i (X_i - 1)^2/2\}}{(1/\sqrt{2\pi})^n \exp\{-\sum_i X_i^2/2\}} > \xi$$
 - algebra: reject H_0 if: $\sum_i X_i > \xi'$
- Find ξ' such that

$$P\left(\sum_{i=1}^n X_i > \xi'; H_0\right) = \alpha$$

- use normal tables

Example (test on normal variance)

- n data points, i.i.d.
 $H_0: X_i \sim N(0, 1)$
 $H_1: X_i \sim N(0, 4)$
- Likelihood ratio test; rejection region:

$$\frac{(1/2\sqrt{2\pi})^n \exp\{-\sum_i X_i^2/(2 \cdot 4)\}}{(1/\sqrt{2\pi})^n \exp\{-\sum_i X_i^2/2\}} > \xi$$
 - algebra: reject H_0 if $\sum_i X_i^2 > \xi'$
- Find ξ' such that

$$P\left(\sum_{i=1}^n X_i^2 > \xi'; H_0\right) = \alpha$$
 - the distribution of $\sum_i X_i^2$ is known
 (derived distribution problem)
 - “chi-square” distribution;
 tables are available

Composite hypotheses

- Got $S = 472$ heads in $n = 1000$ tosses; is the coin fair?
 - $H_0 : p = 1/2$ versus $H_1 : p \neq 1/2$
- Pick a “**statistic**” (e.g., S)
- Pick shape of **rejection region** (e.g., $|S - n/2| > \xi$)
- Choose **significance level** (e.g., $\alpha = 0.05$)
- Pick **critical value** ξ so that:

$$P(\text{reject } H_0; H_0) = \alpha$$

Using the CLT:

$$P(|S - 500| \leq 31; H_0) \approx 0.95; \quad \xi = 31$$

- In our example: $|S - 500| = 28 < \xi$
 H_0 **not rejected** (at the 5% level)

Is my die fair?

- Hypothesis H_0 :
 $P(X = i) = p_i = 1/6, i = 1, \dots, 6$
- Observed occurrences of i : N_i
- Choose form of rejection region;
chi-square test:

$$\text{reject } H_0 \text{ if } T = \sum_i \frac{(N_i - np_i)^2}{np_i} > \xi$$

- Choose ξ so that:

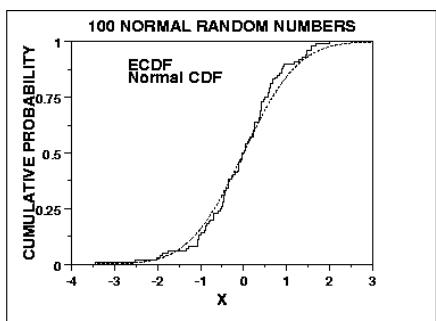
$$P(\text{reject } H_0; H_0) = 0.05$$

$$P(T > \xi; H_0) = 0.05$$

- Need the distribution of T :
(CLT + derived distribution problem)
 - for large n , T has approximately a chi-square distribution
 - available in tables

Do I have the correct pdf?

- Partition the range into bins
 - np_i : expected incidence of bin i (from the pdf)
 - N_i : observed incidence of bin i
 - Use chi-square test (as in die problem)
- Kolmogorov-Smirnov test:
form **empirical CDF**, \hat{F}_X , from data



(<http://www.itl.nist.gov/div898/handbook/>)

- $D_n = \max_x |F_X(x) - \hat{F}_X(x)|$
- $P(\sqrt{n}D_n \geq 1.36) \approx 0.05$

What else is there?

- Systematic methods for coming up with shape of rejection regions
- Methods to estimate an unknown PDF (e.g., form a histogram and “smooth” it out)
- Efficient and recursive signal processing
- Methods to select between less or more complex models
 - (e.g., identify relevant “explanatory variables” in regression models)
- Methods tailored to high-dimensional unknown parameter vectors and huge number of data points (data mining)
- etc. etc....

MIT OpenCourseWare
<http://ocw.mit.edu>

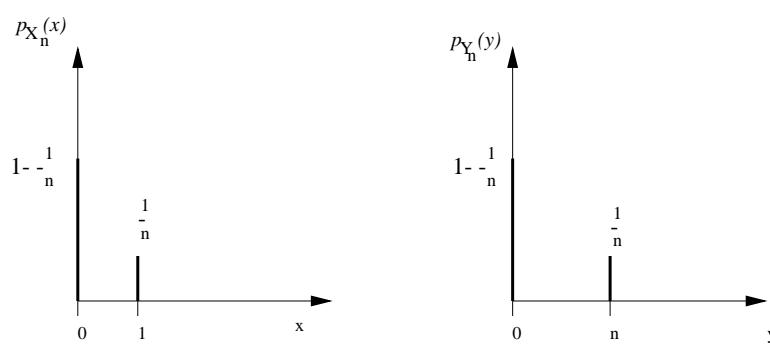
6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Recitation 20: November 18, 2010

1. In your summer internship, you are working for the world's largest producer of lightbulbs. Your manager asks you to estimate the quality of production, that is, to estimate the probability p that a bulb produced by the factory is defectless. You are told to assume that all lightbulbs have the same probability of having a defect, and that defects in different lightbulbs are independent.
 - (a) Suppose that you test n randomly picked bulbs, what is a good estimate Z_n for p , such that Z_n converges to p in probability?
 - (b) If you test 50 light bulbs, what is the probability that your estimate is in the range $p \pm 0.1$?
 - (c) The management asks that your estimate falls in the range $p \pm 0.1$ with probability 0.95. How many light bulbs do you need to test to meet this specification?
- 2.



Let X_n and Y_n have the distributions shown above.

- (a) Find the expected value and variance of X_n and Y_n .
- (b) What does the Chebyshev inequality tell us about the convergence of X_n ? Y_n ?
- (c) Is Y_n convergent in probability? If so, to what value?
- (d) If a sequence of random variables converges in probability to a , does the corresponding sequence of expected values converge to a ? Prove or give a counter example.

A sequence of random variables is said to converge to a number c in the **mean square**, if

$$\lim_{n \rightarrow \infty} \mathbf{E} [(X_n - c)^2] = 0.$$

- (e) Use Markov's inequality to show that convergence in the mean square implies convergence in probability.
- (f) Give an example that shows that convergence in probability does not imply convergence in the mean square.
3. Random variable X is uniformly distributed between -1.0 and 1.0 . Let X_1, X_2, \dots be independent identically distributed random variables with the same distribution as X . Determine which, if any, of the following sequences (all with $i = 1, 2, \dots$) are convergent in probability. Give reasons for your answers. Include the limits if they exist.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Fall 2010)

- (a) X_i
- (b) $Y_i = \frac{X_i}{i}$
- (c) $Z_i = (X_i)^i$

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Recitation 20 Solutions: November 18, 2010

1. (a) Let X_i be a random variable indicating the quality of the i th bulb (“1” for good bulbs, “0” for bad ones). X_i ’s are independent Bernoulli random variables. Let Z_n be

$$Z_n = \frac{X_1 + X_2 + \dots + X_n}{n}.$$

$$\mathbf{E}[Z_n] = p \quad \text{var}(Z_n) = \frac{n\text{var}(X_i)}{n^2} = \frac{\sigma^2}{n},$$

where σ^2 is the variance of X_i .

Applying Chebyshev’s inequality yields,

$$\mathbf{P}(|Z_n - p| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2},$$

As $n \rightarrow \infty$, $\frac{\sigma^2}{n\epsilon^2} \rightarrow 0$ and $\mathbf{P}(|Z_n - p| \geq \epsilon) \rightarrow 0$.
 Hence, Z_n converges to p in probability.

- (b) By Chebychev’s inequality,

$$\mathbf{P}(|Z_{50} - p| \geq 0.1) \leq \frac{\sigma^2}{50(0.1)^2},$$

Since X_i is a Bernoulli random variable, its variance σ^2 is $p(1-p)$, which is less than or equal to $\frac{1}{4}$. Thus,

$$\mathbf{P}(|Z_{50} - p| \geq 0.1) \leq \frac{1/4}{50(0.1)^2} = 0.5$$

- (c) By Chebychev’s inequality,

$$\mathbf{P}(|Z_n - p| \geq 0.1) \leq \frac{\sigma^2}{n\epsilon^2} \leq \frac{1/4}{n(0.1)^2}$$

To guarantee a probability 0.95 of falling in the desired range,

$$\frac{1/4}{n(0.1)^2} < 0.05,$$

which yields $n \geq 500$. Note that $n \geq 500$ guarantees the accuracy specification even for the highest variance, namely $1/4$. For smaller variances, we need smaller values of n to guarantee the desired accuracy. For example, if $\sigma^2 = 1/16$, $n \geq 125$ would suffice.

2. (a) $\mathbf{E}[X_n] = 0 \cdot \left(1 - \frac{1}{n}\right) + 1 \cdot \frac{1}{n} = \frac{1}{n}$
 $\text{var}(X_n) = \left(0 - \frac{1}{n}\right)^2 \cdot \left(1 - \frac{1}{n}\right) + \left(1 - \frac{1}{n}\right)^2 \cdot \left(\frac{1}{n}\right) = \frac{n-1}{n^2}$
 $\mathbf{E}[Y_n] = 0 \cdot \left(1 - \frac{1}{n}\right) + n \cdot \frac{1}{n} = 1$
 $\text{var}(Y_n) = (0 - 1)^2 \cdot \left(1 - \frac{1}{n}\right) + (n - 1)^2 \cdot \left(\frac{1}{n}\right) = n - 1$

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
 (Fall 2010)

(b) Using Chebyshev's inequality, we have

$$\lim_{n \rightarrow \infty} \mathbf{P}\left(\left|X_n - \frac{1}{n}\right| \geq \epsilon\right) \leq \lim_{n \rightarrow \infty} \frac{n-1}{n^2 \epsilon^2} = 0$$

$$\text{Moreover, } \lim_{n \rightarrow \infty} \frac{1}{n} = 0.$$

It follows that X_n converges to 0 in probability. For Y_n , Chebyshev suggests that,

$$\lim_{n \rightarrow \infty} \mathbf{P}\left(\left|Y_n - 1\right| \geq \epsilon\right) \leq \lim_{n \rightarrow \infty} \frac{n-1}{\epsilon^2} = \infty,$$

Thus, we cannot conclude anything about the convergence of Y_n through Chebychev's inequality.

(c) For every $\epsilon > 0$,

$$\lim_{n \rightarrow \infty} \mathbf{P}\left(\left|Y_n\right| \geq \epsilon\right) \leq \lim_{n \rightarrow \infty} \frac{1}{n} = 0,$$

Thus, Y_n converges to zero in probability.

(d) The statement is false. A counter example is Y_n . It converges in probability to 0 yet its expected value is 1 for all n .

(e) Using the Markov bound, we have

$$\mathbf{P}\left(\left|X_n - c\right| \geq \epsilon\right) = P\left(\left|X_n - c\right|^2 \geq \epsilon^2\right) \leq \frac{\mathbf{E}\left[\left(X_n - c\right)^2\right]}{\epsilon^2}.$$

Taking the limit as $n \rightarrow \infty$, we obtain

$$\lim_{n \rightarrow \infty} \mathbf{P}\left(\left|X_n - c\right| \geq \epsilon\right) = 0,$$

which establishes convergence in probability.

(f) A counter example is Y_n . Y_n converges to 0 in probability, but

$$\mathbf{E}\left[\left(Y_n - 0\right)^2\right] = 0 \cdot \left(1 - \frac{1}{n}\right) + (n^2) \cdot \frac{1}{n} = n$$

Thus,

$$\lim_{n \rightarrow \infty} \mathbf{E}\left[\left(Y_n - 0\right)^2\right] = \infty,$$

and Y_n does not converge to 0 in the mean square.

3. (a) No. Since X_i for any $i \geq 1$ is uniformly distributed between -1.0 and 1.0.

(b) Yes, to 0. Since for $\epsilon > 0$,

$$\begin{aligned} \lim_{i \rightarrow \infty} \mathbf{P}\left(\left|Y_i - 0\right| > \epsilon\right) &= \lim_{i \rightarrow \infty} \mathbf{P}\left(\left|\frac{X_i}{i} - 0\right| > \epsilon\right) \\ &= \lim_{i \rightarrow \infty} [\mathbf{P}(X_i > i\epsilon) + \mathbf{P}(X_i < -i\epsilon)] = 0. \end{aligned}$$

(c) Yes, to 0. Since for $\epsilon > 0$,

$$\begin{aligned} \lim_{i \rightarrow \infty} \mathbf{P}\left(\left|Z_i - 0\right| > \epsilon\right) &= \lim_{i \rightarrow \infty} \mathbf{P}\left(\left|(X_i)^i - 0\right| > \epsilon\right) \\ &= \lim_{i \rightarrow \infty} [\mathbf{P}(X_i > \epsilon^{1/i}) + \mathbf{P}(X_i < -(\epsilon)^{1/i})] \\ &= \lim_{i \rightarrow \infty} \left[\frac{1}{2}(1 - \epsilon^{1/i}) + \frac{1}{2}(1 - \epsilon^{1/i})\right] = \lim_{i \rightarrow \infty} (1 - \sqrt[i]{\epsilon}) \\ &= 0. \end{aligned}$$

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Recitation 21
November 23, 2010

1. Let X_1, \dots, X_{10} be independent random variables, uniformly distributed over the unit interval $[0,1]$.

- (a) Estimate $\mathbf{P}(X_1 + \dots + X_{10} \geq 7)$ using the Markov inequality.
- (b) Repeat part (a) using the Chebyshev inequality.
- (c) Repeat part (a) using the central limit theorem.

2. **Problem 10 in the textbook (page 290)**

A factory produces X_n gadgets on day n , where the X_n are independent and identically distributed random variables, with mean 5 and variance 9.

- (a) Find an approximation to the probability that the total number of gadgets produced in 100 days is less than 440.
- (b) Find (approximately) the largest value of n such that

$$\mathbf{P}(X_1 + \dots + X_n \geq 200 + 5n) \leq 0.05.$$

- (c) Let N be the first day on which the total number of gadgets produced exceeds 1000. Calculate an approximation to the probability that $N \geq 220$.

3. Let X_1, X_2, \dots , be independent Poisson random variables with mean and variance equal to 1. For any $n > 0$, let $S_n = \sum_{i=1}^n X_i$.

- (a) Show that S_n is Poisson with mean and variance equal to n . Hint: Relate X_1, X_2, \dots, X_n to a Poisson process with rate 1.
- (b) Show how the central limit theorem suggests the approximation

$$n! \approx \sqrt{2\pi n} \left(\frac{n}{e}\right)^n$$

for large values of the positive integer n .

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Recitation 21 Solutions
November 23, 2010

1. (a) To use the Markov inequality, let $X = \sum_{i=1}^{10} X_i$. Then,

$$\mathbf{E}[X] = 10\mathbf{E}[X_i] = 5,$$

and the Markov inequality yields

$$\mathbf{P}(X \geq 7) \leq \frac{5}{7} = 0.7142.$$

- (b) Using the Chebyshev inequality, we find that

$$\begin{aligned} 2\mathbf{P}(X - 5 \geq 2) &= \mathbf{P}(|X - 5| \geq 2) \\ &\leq \frac{\text{var}(X)}{4} = \frac{10/12}{4} \\ \mathbf{P}(X - 5 \geq 2) &\leq \frac{5}{48} = 0.1042. \end{aligned}$$

- (c) Finally, using the Central Limit Theorem, we find that

$$\begin{aligned} \mathbf{P}\left(\sum_{i=1}^{10} X_i \geq 7\right) &= 1 - \mathbf{P}\left(\sum_{i=1}^{10} X_i \leq 7\right) \\ &= 1 - \mathbf{P}\left(\frac{\sum_{i=1}^{10} X_i - 5}{\sqrt{10/12}} \leq \frac{7 - 5}{\sqrt{10/12}}\right) \\ &\approx 1 - \Phi(2.19) \\ &\approx 0.0143. \end{aligned}$$

2. Check online solutions.
3. (a) If we interpret X_i as the number of arrivals in an interval of length 1 in a Poisson process of rate 1, then, $S_n = X_1 + \dots + X_n$ can be seen as the number of arrivals in an interval of length n in the Poisson process of rate 1. Therefore, S_n is a Poisson random variable with mean and variance equal to n .
- (b) We use the random variables X_1, \dots, X_n and the random variable $S_n = X_1 + \dots + X_n$. Denoting by Z the standard normal, and applying the central limit theorem, we have for

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
 (Fall 2010)

large n

$$\begin{aligned}
 \mathbf{P}(S_n = n) &= \mathbf{P}(n - 1/2 < S_n < n + 1/2) \\
 &= \mathbf{P}\left(\frac{-1}{2\sqrt{n}} < \frac{S_n - n}{\sqrt{n}} \leq \frac{1}{2\sqrt{n}}\right) \\
 &\approx \mathbf{P}\left(\frac{-1}{2\sqrt{n}} < Z \leq \frac{1}{2\sqrt{n}}\right) \\
 &= \frac{1}{\sqrt{2\pi}} \int_{-1/2\sqrt{n}}^{1/2\sqrt{n}} e^{-z^2/2} dz \\
 &\approx \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{n}} e^{-z^2/2} \Big|_{z=0} \\
 &= \frac{1}{\sqrt{2\pi n}}
 \end{aligned}$$

where the first equation follows from the fact that S_n takes integer values, the first approximation is suggested by the central limit theorem, and the second approximation uses the fundamental theorem of calculus (the value of a definite integral over a small interval is equal to the length of the interval times the integrand evaluated at some point within the interval). Since S_n is Poisson with mean n , we have

$$\mathbf{P}(S_n = n) = e^{-n} \frac{n^n}{n!},$$

and by combining the preceding relations, we see that $n! \approx n^n e^{-n} \sqrt{2\pi n} = \sqrt{2\pi n} \left(\frac{n}{e}\right)^n$.

One may show that

$$\lim_{n \rightarrow \infty} \frac{n!}{n^n e^{-n} \sqrt{2\pi n}} = 1,$$

so the relative error of the approximation tends to 0 as $n \rightarrow \infty$. A more precise estimate is that

$$n! = n^n e^{-n} \sqrt{2\pi n} \cdot e^{\lambda_n},$$

where

$$\frac{1}{12n+1} < \lambda_n < \frac{1}{12n}.$$

However, one cannot derive these relations from the central limit theorem.

Note that the form of the approximation was first discovered by de Moivre in the form $n! \approx n^{n+1/2} e^{-n} \cdot (\text{constant})$, and gave a complicated expression for the constant. De Moivre's friend Stirling subsequently showed that the constant has the simple form $\sqrt{2\pi}$.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Recitation 22
November 30, 2010

Examples 8.2, 8.7, 8.12, and 8.15 in the textbook

Romeo and Juliet start dating, but Juliet will be late on any date by a random amount X , uniformly distributed over the interval $[0, \theta]$. The parameter θ is unknown and is modeled as the value of a random variable Θ , uniformly distributed between zero and one hour.

- (a) Assuming that Juliet was late by an amount x on their first date, how should Romeo use this information to update the distribution of Θ ?
- (b) How should Romeo update the distribution of Θ if he observes that Juliet is late by x_1, \dots, x_n on the first n dates? Assume that Juliet is late by a random amount X_1, \dots, X_n on the first n dates where, given θ , X_1, \dots, X_n are uniformly distributed between zero and θ and are conditionally independent.
- (c) Find the MAP estimate of Θ based on the observation $X = x$.
- (d) Find the LMS estimate of Θ based on the observation $X = x$.
- (e) Calculate the conditional mean squared error for the MAP and the LMS estimates. Compare your results.
- (f) Derive the linear LMS estimator of Θ based on X .
- (g) Calculate the conditional mean squared error for the linear LMS estimate. Compare your answer to the results of part (e).

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Recitation 22 Solutions

The problem is based on the following examples in the textbook.

- (a) Example 8.2 page 414.
- (b) Example 8.2 page 414.
- (c) Example 8.7 page 424.
- (d) Example 8.7 page 424.
- (e) Example 8.12 page 432-433.
- (f) Example 8.15 page 439-440.
- (g) Example 8.15 page 439-440.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Recitation 23
December 2, 2010

1. Example 9.1, page 463 in textbook

Romeo and Juliet start dating, but Juliet will be late on any date by a random amount X , uniformly distributed over the interval $[0, \theta]$. The parameter θ is unknown. Assuming that Juliet was late by an amount x on their first date, find the ML estimate of θ based on the observation $X = x$.

2. Example 9.4, page 464 in textbook

Estimate the mean μ and variance v of a normal distribution using n independent observations X_1, \dots, X_n .

3. Example 9.8, page 474 of textbook

We would like to estimate the fraction of voters supporting a particular candidate for office. We collect n independent sample voter responses X_1, \dots, X_n , where X_i is viewed as a Bernoulli random variable, with $X_i = 1$ if the i th voter supports the candidate. We conducted a poll of 1200 people in North Carolina, and found that 684 were supporting the candidate. We would like to construct a 95% confidence interval for θ , the proportion of people who support the candidate. As we saw in lecture, using the central limit theorem, an (approximate) 95% confidence interval can be defined as

$$\hat{\Theta}^- = \hat{\Theta}_n - 1.96\sqrt{\frac{v}{n}}, \quad \hat{\Theta}^+ = \hat{\Theta}_n + 1.96\sqrt{\frac{v}{n}}$$

where $v = \text{Var}(X_i)$, and $\hat{\Theta}_n = (X_1 + \dots + X_n)/n$. Unfortunately, we don't know the value for v . Construct confidence intervals for θ using the following three ways of estimating or bounding the value for v (in each case simply assume that v is equal to the given estimate; note that this is a further approximation in cases (a) and (b)).

(a)

$$\hat{S}_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \hat{\Theta}_n)^2$$

(b)

$$\hat{\Theta}_n(1 - \hat{\Theta}_n)$$

(c) The most conservative upper bound for the variance.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Recitation 23 Solutions

1. Example 9.1 in the textbook (page 463)
2. Example 9.4 in the textbook (page 464)
3. Example 9.8 in the textbook (page 474)

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Recitation 24
December 7, 2010

1. A blackbody at temperature θ radiates photons of all wavelengths, described by its characteristic spectrum. This problem will have you estimate θ , which is fixed but unknown. The PMF for the number of photons K in a given wavelength range and a fixed very short time interval is given by,

$$p_K(k; \theta) = \frac{1}{Z(\theta)} e^{-k/\theta}, k = 0, 1, 2, \dots$$

$Z(\theta)$ is a normalization factor for the probability distribution (the physicists call it the partition function). You are given the task of determining the temperature of the body to two significant digits by photon counting in non-overlapping time intervals of duration one second. The photon emissions in non-overlapping time intervals are statistically independent from each other.

- (a) Determine the normalization factor $Z(\theta)$.
- (b) Compute the expected value of the photon number measured in any 1 second time interval, $\mu_K = \mathbf{E}_\theta[K]$, and its variance, $\text{var}_\theta(K) = \sigma_K^2$.
- (c) You count the number k_i of photons detected in n non-overlapping 1 second time intervals. Find the maximum likelihood estimator, $\hat{\theta}_n$, for temperature θ . Note, it might be useful to introduce the average photon number $s_n = \frac{1}{n} \sum_{i=1}^n k_i$. In order to keep the analysis simple we assume that the body is hot, i.e. $\theta \gg 1$. You may use the approximation: $\frac{1}{e^{1/\theta}-1} \approx \theta$ for $\theta \gg 1$.

In the following questions we wish to estimate the mean of the photon count in a one second time interval using the estimator \hat{K} , which is given by,

$$\hat{K} = \frac{1}{n} \sum_{i=1}^n K_i.$$

- (d) Find the number of samples n for which the noise to signal ratio for \hat{K} , (i.e., $\frac{\sigma_{\hat{K}}}{\mu_{\hat{K}}}$), is 0.01.
 - (e) Find a 95% confidence interval for the mean photon count estimate for the situation in part (d). (You may use the central limit theorem.)
2. Given the five data pairs (x_i, y_i) in the table below,

x	0.8	2.5	5	7.3	9.1
y	-2.3	20.9	103.5	215.8	334

we want to construct a model relating x and y . We consider a linear model

$$Y_i = \theta_0 + \theta_1 x_i + W_i, \quad i = 1, \dots, 5,$$

and a quadratic model

$$Y_i = \beta_0 + \beta_1 x_i^2 + V_i, \quad i = 1, \dots, 5.$$

where W_i and V_i represent additive noise terms, modeled by independent normal random variables with mean zero and variance σ_1^2 and σ_2^2 , respectively.

- (a) Find the ML estimates of the linear model parameters.
- (b) Find the ML estimates of the quadratic model parameters.

Note: You may use the regression formulas and the connection with ML described in pages 478-479 of the text. However, the regression material is outside the scope of the final.

The figure below shows the data points (x_i, y_i) , $i = 1, \dots, 5$, the estimated linear model

$$y = 40.53x - 65.86,$$

and the estimated quadratic model

$$y = 4.09x^2 - 3.07.$$

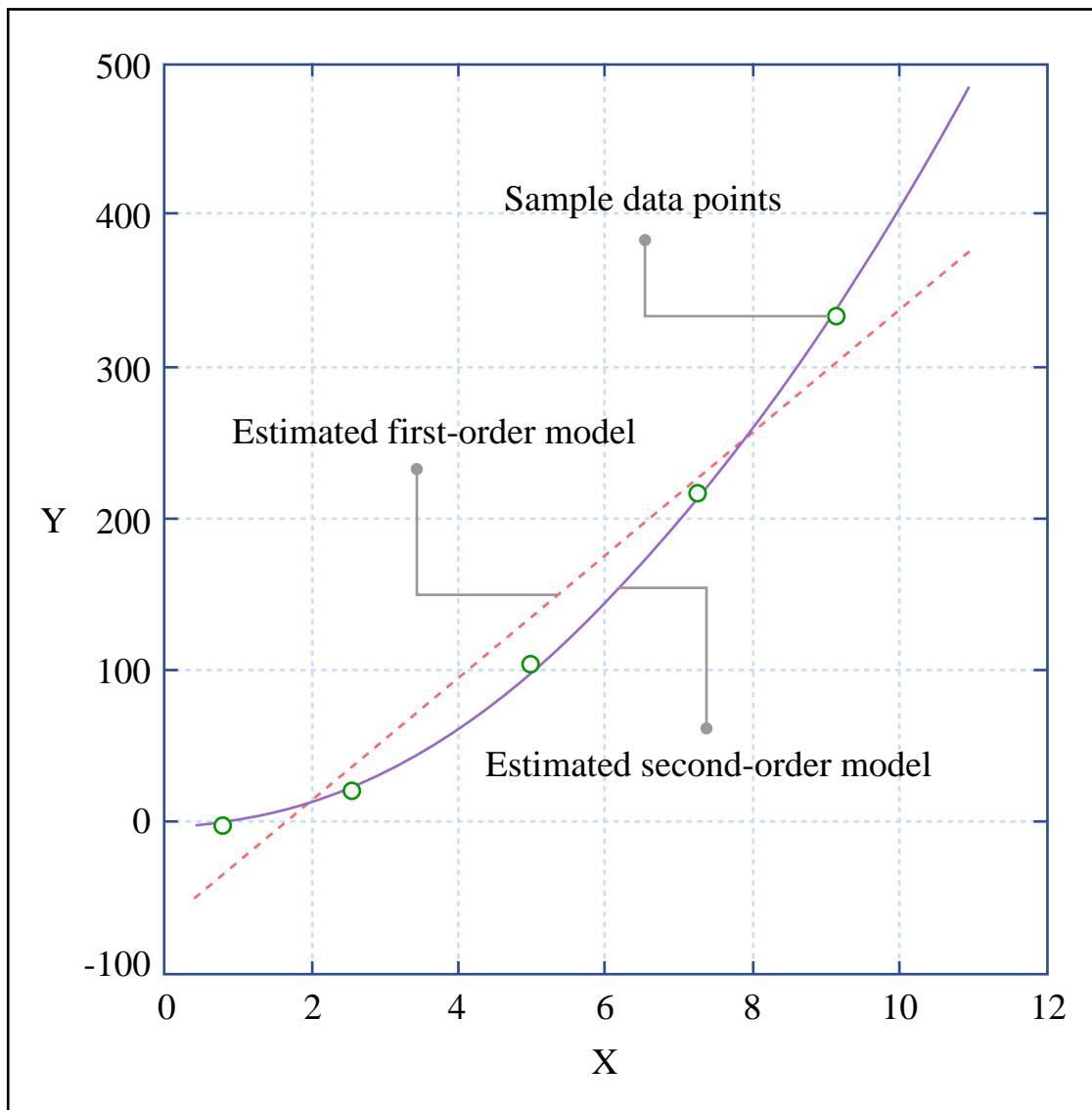


Image by MIT OpenCourseWare.

Figure 1: Regression Plot

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Recitation 24: Solutions
December 7, 2010

1. (a) Normalization of the distribution requires:

$$1 = \sum_{k=0}^{\infty} p_K(k; \theta) = \sum_{k=0}^{\infty} \frac{e^{-k/\theta}}{Z(\theta)} = \frac{1}{Z(\theta)} \sum_{k=0}^{\infty} e^{-k/\theta} = \frac{1}{Z(\theta) \cdot (1 - e^{-1/\theta})},$$

$$\text{so } Z(\theta) = \frac{1}{1 - e^{-1/\theta}}.$$

- (b) Rewriting $p_K(k; \theta)$ as:

$$p_K(k; \theta) = \left(e^{-1/\theta} \right)^k \left(1 - e^{-1/\theta} \right), \quad k = 0, 1, \dots$$

the probability distribution for the photon number is a geometric probability distribution with probability of success $p = 1 - e^{-1/\theta}$, and it is shifted with 1 to the left since it starts with $k = 0$. Therefore the photon number expectation value is

$$\mu_K = \frac{1}{p} - 1 = \frac{1}{1 - e^{-1/\theta}} - 1 = \frac{1}{e^{1/\theta} - 1}$$

and its variance is

$$\sigma_K^2 = \frac{1-p}{p^2} = \frac{e^{-1/\theta}}{(1 - e^{-1/\theta})^2} = \mu_K^2 + \mu_K.$$

- (c) The joint probability distribution for the k_i is

$$p_K(k_1, \dots, k_n; \theta) = \frac{1}{Z(\theta)^n} \prod_{i=1}^n e^{-k_i/\theta} = \frac{1}{Z(\theta)^n} e^{-\frac{1}{\theta} \sum_{i=1}^n k_i}.$$

The log likelihood is $-n \cdot \log Z(\theta) - 1/\theta \sum_{i=1}^n k_i$.

We find the maxima of the log likelihood by setting the derivative with respect to the parameter θ to zero:

$$\frac{d}{d\theta} \log p_K(k_1, \dots, k_n; \theta) = -n \cdot \frac{e^{-1/\theta}}{\theta^2 (1 - e^{-1/\theta})} + \frac{1}{\theta^2} \sum_{i=1}^n k_i = 0$$

or

$$\frac{1}{e^{1/\theta} - 1} = \frac{1}{n} \sum_{i=1}^n k_i = s_n.$$

For a hot body, $\theta \gg 1$ and $\frac{1}{e^{1/\theta} - 1} \approx \theta$, we obtain

$$\theta \approx \frac{1}{n} \sum_{i=1}^n k_i = s_n.$$

Thus the maximum likelihood estimator $\hat{\Theta}_n$ for the temperature is given in this limit by the sample mean of the photon number

$$\hat{\Theta}_n = \frac{1}{n} \sum_{i=1}^n K_i.$$

- (d) According to the central limit theorem, the sample mean for large enough n (in the limit) approaches a Gaussian distribution with standard deviation our root mean square error

$$\sigma_{\hat{\Theta}_n} = \frac{\sigma_K}{\sqrt{n}}.$$

To allow only for 1% relative root mean square error in the temperature, we need $\frac{\sigma_K}{\sqrt{n}} < 0.01\mu_K$. With $\sigma_K^2 = \mu_K^2 + \mu_K$ it follows that

$$\sqrt{n} > \frac{\sigma_K}{0.01\mu_K} = 100 \frac{\sqrt{\mu_K^2 + \mu_K}}{\mu_K} = 100 \sqrt{1 + \frac{1}{\mu_K}}.$$

In general, for large temperatures, i.e. large mean photon numbers $\mu_K \gg 1$, we need about 10,000 samples.

- (e) The 95% confidence interval for the temperature estimate for the situation in part (d), i.e.

$$\sigma_{\hat{\Theta}_n} = \frac{\sigma_K}{\sqrt{n}} = 0.01\mu_K,$$

is

$$[\hat{K} - 1.96\sigma_{\hat{K}}, \hat{K} + 1.96\sigma_{\hat{K}}] = [\hat{K} - 0.0196\mu_K, \hat{K} + 0.0196\mu_K].$$

2. (a) Using the regression formulas of Section 9.2, we have

$$\hat{\theta}_1 = \frac{\sum_{i=1}^5 (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^5 (x_i - \bar{x})^2}, \quad \hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x},$$

where

$$\bar{x} = \frac{1}{5} \sum_{i=1}^5 x_i = 4.94, \quad \bar{y} = \frac{1}{5} \sum_{i=1}^5 y_i = 134.38.$$

The resulting ML estimates are

$$\hat{\theta}_1 = 40.53, \quad \hat{\theta}_0 = -65.86.$$

- (b) Using the same procedure as in part (a), we obtain

$$\hat{\theta}_1 = \frac{\sum_{i=1}^5 (x_i^2 - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^5 (x_i^2 - \bar{x})^2}, \quad \hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x},$$

where

$$\bar{x} = \frac{1}{5} \sum_{i=1}^5 x_i^2 = 33.60, \quad \bar{y} = \frac{1}{5} \sum_{i=1}^5 y_i = 134.38.$$

which for the given data yields

$$\hat{\theta}_1 = 4.09, \quad \hat{\theta}_0 = -3.07.$$

Figure 1 shows the data points (x_i, y_i) , $i = 1, \dots, 5$, the estimated linear model

$$y = 40.53x - 65.86,$$

and the estimated quadratic model

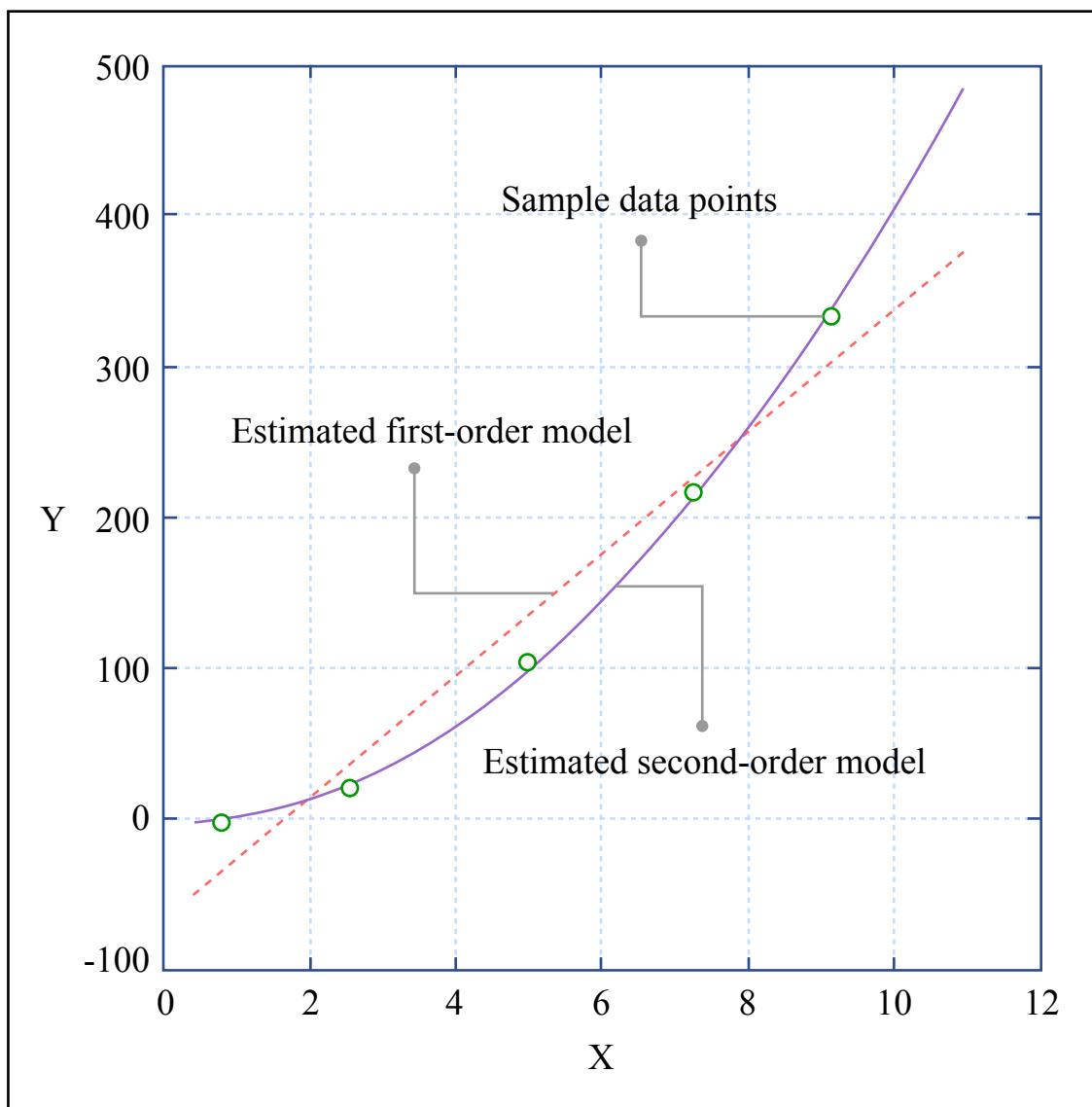


Image by MIT OpenCourseWare.

Figure 1: Regression Plot

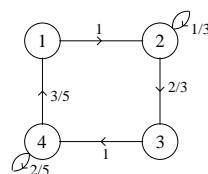
MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Tutorial 10
November 18/19, 2010

1. Define X as the height in meters of a randomly selected Canadian, where the selection probability is equal for each Canadian, and denote $\mathbf{E}[X]$ by h . Bo is interested in estimating h . Because he is sure that no Canadian is taller than 3 meters, Bo decides to use 1.5 meters as a conservative (large) value for the standard deviation of X . To estimate h , Bo averages the heights of n Canadians that he selects at random; he denotes this quantity by H .
 - (a) In terms of h and Bo's 1.5 meter bound for the standard deviation of X , determine the expected value and standard deviation for H .
 - (b) Help Bo by calculating a minimum value of n (with $n > 0$) such that the standard deviation of Bo's estimator, H , will be less than 0.01 meters.
 - (c) Bo would like to be 99% sure that his estimate is within 5 centimeters of the true average height of Canadians. Using the Chebyshev inequality, calculate the minimum value of n that will make Bo happy.
 - (d) If we agree that no Canadians are taller than three meters, why is it correct to use 1.5 meters as an upper bound on the standard deviation for X , the height of any Canadian selected at random?
2. On any given week while taking 6.041, a student can be either up-to-date on learning the material, or she may have fallen behind. If she is up-to-date in a given week, the probability that she will be up-to-date (or behind) in the next week is 0.8 (or 0.2, respectively). If she is behind in the given week, the probability that she will be up-to-date (or behind) in the next week is 0.6 (or 0.4, respectively). We assume that these probabilities do not depend on whether she was up-to-date or behind in previous weeks, so we can model the situation as a 2-state Markov chain where State 1 is the case when the student is up-to-date and State 2 is the case when the student is behind.
 - (a) Calculate the mean first passage time to State 1, starting from State 2.
 - (b) Calculate the mean recurrence time to State 1.
3. Consider the following Markov chain:



The steady-state probabilities for this process are:

$$\pi_1 = \frac{6}{31} \quad \pi_2 = \frac{9}{31} \quad \pi_3 = \frac{6}{31} \quad \pi_4 = \frac{10}{31}$$

Assume the process is in state 1 just before the first transition.

- (a) Determine the expected value and variance of K , the number of transitions up to and including the next transition on which the process returns to state 1.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Fall 2010)

- (b) What is the probability that the state of the system resulting from transition 1000 is neither the same as the state resulting from transition 999 nor the same as the state resulting from transition 1001?

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Tutorial 10 Solutions
November 18/19, 2010

1. Note that n is deterministic and H is a random variable.

(a) Use X_1, X_2, \dots to denote the (random) measured heights.

$$\begin{aligned} H &= \frac{X_1 + X_2 + \cdots + X_n}{n} \\ \mathbf{E}[H] &= \frac{\mathbf{E}[X_1 + X_2 + \cdots + X_n]}{n} = \frac{n\mathbf{E}[X]}{n} = h \\ \sigma_H &= \sqrt{\text{var}(H)} = \sqrt{\frac{n \text{var}(X)}{n^2}} \quad (\text{var of sum of independent r.v.s is sum of vars}) \\ &= \frac{1.5}{\sqrt{n}} \end{aligned}$$

(b) We solve $\frac{1.5}{\sqrt{n}} < 0.01$ for n to obtain $n > 22500$.

(c) Apply the Chebyshev inequality to H with $\mathbf{E}[H]$ and $\text{var}(H)$ from part (a):

$$\begin{aligned} \mathbf{P}(|H - h| \geq t) &\leq \left(\frac{\sigma_H}{t}\right)^2 \\ \mathbf{P}(|H - h| < t) &\geq 1 - \left(\frac{\sigma_H}{t}\right)^2 \end{aligned}$$

To be “99% sure” we require the latter probability to be at least 0.99. Thus we solve

$$1 - \left(\frac{\sigma_H}{t}\right)^2 \geq 0.99$$

with $t = 0.05$ and $\sigma_H = \frac{1.5}{\sqrt{n}}$ to obtain

$$n \geq \left(\frac{1.5}{0.05}\right)^2 \frac{1}{0.01} = 90000.$$

(d) Intuitively, the variance of a random variable X that takes values in the range $[0, b]$ is maximum when X takes the value 0 with probability 0.5 and the value b with probability 0.5, in which case the variance of X is $b^2/4$ and its standard deviation is $b/2$.

More formally, since $\mathbf{E}[(X - c)^2]$ is minimized when $c = \mathbf{E}[X]$, we have for any random variable X taking values in $[0, b]$,

$$\begin{aligned} \text{var}(X) &\leq \mathbf{E}[(X - \frac{b}{2})^2] \\ &= \mathbf{E}[X^2] - b\mathbf{E}[X] + \frac{b^2}{4} \\ &= \mathbf{E}[X(X - b)] + \frac{b^2}{4} \\ &\leq 0 + \frac{b^2}{4}, \end{aligned}$$

since $0 \leq X \leq b \Rightarrow X(X - b) \leq 0$. Thus $\sigma_X \leq b/2$.

In our example, we have $b = 3$, so $\sigma_X \leq 3/2$.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
 (Fall 2010)

2. (a) Setting $s = 1$, we get $t_1 = 0$ and

$$\begin{aligned} t_2 &= 1 + \sum_{j=1}^m p_{ij}t_j \quad \forall i \neq s, \\ &= 1 + p_{22}t_2 \\ &\Rightarrow t_2 = 5/3. \end{aligned}$$

(b)

$$\begin{aligned} t_s^* &= 1 + \sum_{j=1}^m p_{sj}t_j \\ t_1^* &= 1 + p_{12}t_2 = 4/3. \end{aligned}$$

3. (a) $K = 2 + X_1 + X_2$, where X_1 and X_2 are independent exponential random variables with parameters $2/3$ and $3/5$.

$$\begin{aligned} E[K] &= 2 + 1/p_1 + 1/p_2 \\ &= 31/6. \\ \text{var}(K) &= \frac{1-p_1}{p_1^2} + \frac{1-p_2}{p_2^2} \\ &= 67/36. \end{aligned}$$

(b)

$$\begin{aligned} \mathbf{P}(A) &= \mathbf{P}(X_{999} \neq X_{1000} \neq X_{1001}) \\ &= \sum_{i=1}^4 \mathbf{P}(A|X_{999} = i)\pi_i \\ &= 2/3\pi_1 + 2/3\pi_2 + 3/5\pi_3 + 3/5\pi_4 \\ &= 30/93 + 48/155 \approx 0.6323. \end{aligned}$$

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

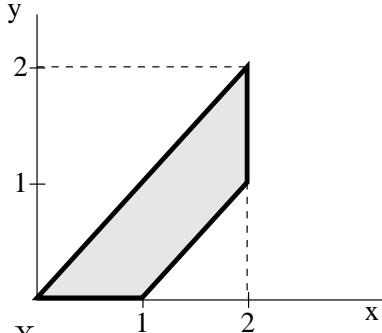
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Tutorial 11

1. Continuous random variables X and Y have a joint PDF given by

$$f_{X,Y}(x,y) = \begin{cases} 2/3 & \text{if } (x,y) \text{ belongs to the closed shaded region} \\ 0 & \text{otherwise} \end{cases}$$



We want to estimate Y based on X .

- (a) Find the LMS estimator $g(X)$ of Y .
 - (b) Calculate the conditional mean squared error $\mathbf{E}[(Y - g(X))^2 | X = x]$.
 - (c) Calculate the mean squared error $\mathbf{E}[(Y - g(X))^2]$. Is it the same as $\mathbf{E}[\text{var}(Y|X)]$?
 - (d) Derive $L(X)$, the linear LMS estimator of Y based on X .
 - (e) How do you expect the mean squared error of $L(X)$ to compare to that of $g(X)$?
 - (f) What problem do you expect to encounter, if any, if you try to find the MAP estimator for Y based on observations of X .
2. Consider a noisy channel over which you send messages consisting of 0s and 1s to your friend. It is known that the channel independently flips each bit sent with some fixed probability p ; however the value of p is unknown. You decide to conduct some experiments to estimate p and seek your friend's help. Your friend, cheeky as she is, insists that you send her messages consisting of three bits each (which you will both agree upon in advance); for each message, she will only tell you the total number of bits in that message that were flipped. Let X denote the number of bits flipped in a particular three-bit message.
- (a) Find the PMF of X .
 - (b) Derive the ML estimator for p based on X_1, \dots, X_n , the numbers of bits flipped in the first n three-bit messages.
 - (c) Is the ML estimator unbiased?
 - (d) Is the ML estimator consistent?
 - (e) You send $n = 100$ three-bit messages and find that the total number of bits flipped is 20. Construct a 95% confidence interval for p . If necessary, you may use a conservative bound on the variance of the number of bits flipped.
 - (f) What are some other ways to estimate the variance. How do you expect your confidence interval to change with different estimates of the variance.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Tutorial 11 Solutions

1. (a) The LMS estimator is

$$g(x) = \mathbf{E}[Y|X] = \begin{cases} \frac{1}{2}X & 0 \leq X < 1 \\ X - \frac{1}{2} & 1 \leq X \leq 2 \\ \text{Undefined} & \text{Otherwise} \end{cases}$$

(b) If $x \in [0, 1]$, the conditional PDF of Y is uniform over the interval $[0, x]$, and

$$\mathbf{E}[(Y - g(X))^2 | X = x] = \frac{x^2}{12}.$$

Similarly, if $x \in [1, 2]$, the conditional PDF of Y is uniform over $[1 - x, x]$, and

$$\mathbf{E}[(Y - g(X))^2 | X = x] = 1/12.$$

(c) The expectations $\mathbf{E}[(Y - g(X))^2]$ and $\mathbf{E}[\text{var}(Y|X)]$ are equal because by the law of iterated expectations,

$$\mathbf{E}[(Y - g(X))^2] = \mathbf{E}[\mathbf{E}[(Y - g(X))^2 | X]] = \mathbf{E}[\text{var}(Y | X)].$$

Recall from part (b) that

$$\text{var}(Y|X = x) = \begin{cases} \frac{x^2}{12} & 0 \leq x < 1, \\ \frac{1}{12} & 1 \leq x \leq 2. \end{cases}$$

It follows that

$$\mathbf{E}[\text{var}(Y | X)] = \int_x \text{var}(Y | X = x) f_X(x) dx = \int_0^1 \frac{x^2}{12} \cdot \frac{2}{3} x dx + \int_1^2 \frac{1}{12} \cdot \frac{2}{3} dx = \frac{5}{72}.$$

Note that

$$f_X(x) = \begin{cases} 2x/3 & 0 \leq x < 1, \\ 2/3 & 1 \leq x \leq 2. \end{cases}$$

- (d) The linear LMS estimator is

$$L(X) = \mathbf{E}[Y] + \frac{\text{cov}(X, Y)}{\text{var}(X)}[X - \mathbf{E}[X]].$$

In order to calculate $\text{var}(X)$ we first calculate $\mathbf{E}[X^2]$ and $\mathbf{E}[X]^2$.

$$\begin{aligned} \mathbf{E}[X^2] &= \int_0^2 x^3 \frac{2}{3} dx + \int_1^2 x^2 \frac{2}{3} dx, \\ &= \frac{31}{18}, \\ \mathbf{E}[X] &= \int_0^2 x^2 \frac{2}{3} dx + \int_1^2 x \frac{2}{3} dx, \\ &= \frac{11}{9} \end{aligned}$$

$$\text{var}(X) = \mathbf{E}[X^2] - \mathbf{E}[X]^2 = \frac{37}{162}.$$

$$\mathbf{E}[Y] = \int_0^1 \int_0^x \frac{2}{3}y \, dy \, dx + \int_1^2 \int_{x-1}^x \frac{2}{3}y \, dy \, dx = \frac{1}{9} + \frac{2}{3} = \frac{7}{9}.$$

To determine $\text{cov}(X, Y)$ we need to evaluate $\mathbf{E}[XY]$.

$$\begin{aligned}\mathbf{E}[XY] &= \int_x \int_y xy f_{X,Y}(x,y) dy dx \\ &= \int_0^1 \int_0^x yx \frac{2}{3} dy dx + \int_1^2 \int_{x-1}^x yx \frac{2}{3} dy dx \\ &= \frac{41}{36}\end{aligned}$$

Therefore $\text{cov}(X, Y) = \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y] = \frac{61}{324}$. Therefore,

$$L(X) = \frac{7}{9} + \frac{61}{74}[X - \frac{11}{9}].$$

- (e) The LMS estimator is the one that minimizes mean squared error (among all estimators of Y based on X). The linear LMS estimator, therefore, cannot perform better than the LMS estimator, i.e., we expect $\mathbf{E}[(Y - L(X))^2] \geq \mathbf{E}[(Y - g(X))^2]$. In fact,

$$\begin{aligned}\mathbf{E}[(Y - L(X))^2] &= \sigma_Y^2(1 - \rho^2), \\ &= \sigma_Y^2\left(1 - \frac{\text{cov}(X, Y)^2}{\sigma_X^2 \sigma_Y^2}\right), \\ &= \frac{37}{162} \left(1 - \left(\frac{61}{74}\right)^2\right), \\ &= 0.073 \geq \frac{5}{72}\end{aligned}$$

- (f) For a single observation x of X , the MAP estimate is not unique since all possible values of Y for this x are equally likely. Therefore, the MAP estimator does not give meaningful results.
2. (a) X is a binomial random variable with parameters $n = 3$ and given the probability p that a single bit is flipped in a transmission over the noisy channel:

$$p_X(k; p) = \begin{cases} \binom{3}{k} p^k (1-p)^{3-k}, & k = 0, 1, 2, 3 \\ 0 & \text{o.w.} \end{cases}$$

- (b) To derive the ML estimator for p based on X_1, \dots, X_n , the numbers of bits flipped in the first n three-bit messages, we need to find the value of p that maximizes the likelihood function:

$$\hat{p}_n = \arg \max_p p_{X_1, \dots, X_n}(k_1, k_2, \dots, k_n; p)$$

Since the X_i 's are independent, the likelihood function simplifies to:

$$p_{X_1, \dots, X_n}(k_1, k_2, \dots, k_n; p) = \prod_{i=1}^n p_{X_i}(k_i; p) = \prod_{i=1}^n \binom{3}{k_i} p^{k_i} (1-p)^{3-k_i}$$

The log-likelihood function is given by

$$\log(p_{X_1, \dots, X_n}(k_1, k_2, \dots, k_n; p)) = \sum_{i=1}^n \left(k_i \log(p) + (3 - k_i) \log(1 - p) + \log \binom{3}{k_i} \right)$$

We then maximize the log-likelihood function with respect to p :

$$\begin{aligned} \frac{1}{p} \left(\sum_{i=1}^n k_i \right) - \frac{1}{1-p} \left(\sum_{i=1}^n (3 - k_i) \right) &= 0 \\ \left(3n - \sum_{i=1}^n k_i \right) p &= \left(\sum_{i=1}^n k_i \right) (1-p) \\ \hat{p}_n &= \frac{1}{3n} \sum_{i=1}^n k_i \end{aligned}$$

This yields the ML estimator:

$$\hat{P}_n = \frac{1}{3n} \sum_{i=1}^n X_i$$

(c) The estimator is unbiased since:

$$\begin{aligned} \mathbf{E}_p[\hat{P}_n] &= \frac{1}{3n} \sum_{i=1}^n \mathbf{E}_p[X_i] \\ &= \frac{1}{3n} \sum_{i=1}^n 3p \\ &= p \end{aligned}$$

- (d) By the weak law of large numbers, $\frac{1}{n} \sum_{i=1}^n X_i$ converges in probability to $\mathbf{E}_p[X_i] = 3p$, and therefore $\hat{P}_n = \frac{1}{3n} \sum_{i=1}^n X_i$ converges in probability to p . Thus \hat{P}_n is consistent.
- (e) Sending 3 bit messages instead of 1 bit messages does not affect the ML estimate of p . To see this, let Y_i be a Bernoulli RV which takes the value 1 if the i th bit is flipped (with probability p), and let $m = 3n$ be the total number of bits sent over the channel. The ML estimate of p is then

$$\hat{P}_n = \frac{1}{3n} \sum_{i=1}^n X_i = \frac{1}{m} \sum_{i=1}^m Y_i.$$

Using the central limit theorem, \hat{P}_n is approximately a normal RV for large n . An approximate 95% confidence interval for p is then,

$$\left[\hat{P}_n - 1.96 \sqrt{\frac{v}{m}}, \hat{P}_n + 1.96 \sqrt{\frac{v}{m}} \right]$$

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Fall 2010)

where v is the variance of Y_i .

As suggested by the question, we estimate the unknown variance v by the conservative upper bound of $1/4$. We are also given that $n = 100$ and the number of bits flipped is 20, yielding $\hat{P}_n = \frac{2}{30}$. Thus, an approximate 95% confidence interval is $[0.01, 0.123]$.

- (f) Other estimates for the variance are the sample variance and the estimate $\hat{P}_n(1 - \hat{P}_n)$. They potentially result in narrower confidence intervals than the conservative variance estimate used in part (e).

MIT OpenCourseWare
<http://ocw.mit.edu>

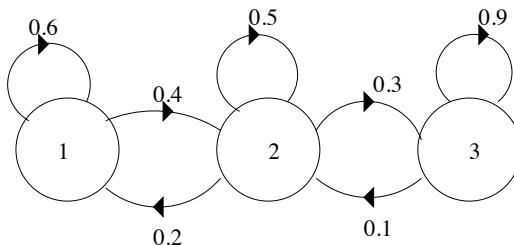
6.041SC Probabilistic Systems Analysis and Applied Probability

Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
 (Final Exam | Fall 2010)

Problem 1. (32 points) Consider a Markov chain $\{X_n; n = 0, 1, \dots\}$, specified by the following transition diagram.



1. (4 points) Given that the chain starts with $X_0 = 1$, find the probability that $X_2 = 2$.
2. (4 points) Find the steady-state probabilities π_1, π_2, π_3 of the different states.

In case you did not do part (b) correctly, in all subsequent parts of this problem you can just use the symbols π_i : you do not need to plug in actual numbers.

3. (4 points) Let $Y_n = X_n - X_{n-1}$. Thus, $Y_n = 1$ indicates that the n th transition was to the right, $Y_n = 0$ indicates it was a self-transition, and $Y_n = -1$ indicates it was a transition to the left. Find $\lim_{n \rightarrow \infty} \mathbf{P}(Y_n = 1)$.
4. (4 points) Is the sequence Y_n a Markov chain? Justify your answer.
5. (4 points) Given that the n th transition was a transition to the right ($Y_n = 1$), find the probability that the previous state was state 1. (You can assume that n is large.)
6. (4 points) Suppose that $X_0 = 1$. Let T be defined as the first *positive time* at which the state is again equal to 1. Show how to find $\mathbf{E}[T]$. (It is enough to write down whatever equation(s) needs to be solved; you do not have to actually solve it/them or to produce a numerical answer.)
7. (4 points) Does the sequence X_1, X_2, X_3, \dots converge in probability? If yes, to what? If not, just say “no” without explanation.
8. (4 points) Let $Z_n = \max\{X_1, \dots, X_n\}$. Does the sequence Z_1, Z_2, Z_3, \dots converge in probability? If yes, to what? If not, just say “no” without explanation.

Problem 2. (68 points) Alice shows up at an Athena* cluster at time zero and spends her time exclusively in typing emails. The times that her emails are sent are a Poisson process with rate λ_A per hour.

1. (3 points) What is the probability that Alice sent exactly three emails during the time interval $[1, 2]$?
2. Let Y_1 and Y_2 be the times at which Alice’s first and second emails were sent.
 - (a) (3 points) Find $\mathbf{E}[Y_2 | Y_1]$.
 - (b) (3 points) Find the PDF of Y_1^2 .
 - (c) (3 points) Find the joint PDF of Y_1 and Y_2 .

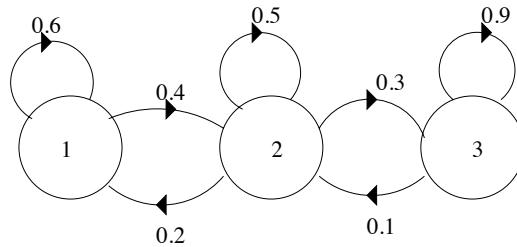
*Athena is MIT’s UNIX-based computing environment. OCW does not provide access to it.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Final Exam | Fall 2010)

3. You show up at time 1 and you are told that Alice has sent exactly one email so far. (Only give answers here, no need to justify them.)
 - (a) **(3 points)** What is the conditional expectation of Y_2 given this information?
 - (b) **(3 points)** What is the conditional expectation of Y_1 given this information?
4. Bob just finished exercising (without email access) and sits next to Alice at time 1. He starts typing emails at time 1, and fires them according to an independent Poisson process with rate λ_B .
 - (a) **(5 points)** What is the PMF of the total number of emails sent by the two of them together during the interval $[0, 2]$?
 - (b) **(5 points)** What is the expected value of the total typing time associated with the email that Alice is typing at the time that Bob shows up? (Here, “total typing time” includes the time that Alice spent on that email both before and after Bob’s arrival.)
 - (c) **(5 points)** What is the expected value of the time until each one of them has sent at least one email? (Note that we count time starting from time 0, and we take into account any emails possibly sent out by Alice during the interval $[0, 1]$.)
 - (d) **(5 points)** Given that a total of 10 emails were sent during the interval $[0, 2]$, what is the probability that exactly 4 of them were sent by Alice?
5. **(5 points)** Suppose that $\lambda_A = 4$. Use Chebyshev’s inequality to find an upper bound on the probability that Alice sent at least 5 emails during the time interval $[0, 1]$. Does the Markov inequality provide a better bound?
6. **(5 points)** You do not know λ_A but you watch Alice for an hour and see that she sent exactly 5 emails. Derive the maximum likelihood estimate of λ_A based on this information.
7. **(5 points)** We have reasons to believe that λ_A is a large number. Let N be the number of emails sent during the interval $[0, 1]$. Justify why the CLT can be applied to N , and give a precise statement of the CLT in this case.
8. **(5 points)** Under the same assumption as in last part, that λ_A is large, you can now pretend that N is a normal random variable. Suppose that you observe the value of N . Give an (approximately) 95% confidence interval for λ_A . State precisely what approximations you are making.
Possibly useful facts: The cumulative normal distribution satisfies $\Phi(1.645) = 0.95$ and $\Phi(1.96) = 0.975$.
9. You are now told that λ_A is actually the realized value of an exponential random variable Λ , with parameter 2:
$$f_\Lambda(\lambda) = 2e^{-2\lambda}, \quad \lambda \geq 0.$$
 - (a) **(5 points)** Find $\mathbf{E}[N^2]$.
 - (b) **(5 points)** Find the linear least squares estimator of Λ given N .

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Final Exam | Fall 2010)

Problem 1. (32 points) Consider a Markov chain $\{X_n; n = 0, 1, \dots\}$, specified by the following transition diagram.



1. (4 points) Given that the chain starts with $X_0 = 1$, find the probability that $X_2 = 2$.

2. (4 points) Find the steady-state probabilities π_1, π_2, π_3 of the different states.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Final Exam | Fall 2010)

In case you did not do part (b) correctly, in all subsequent parts of this problem you can just use the symbols π_i : you do not need to plug in actual numbers.

3. **(4 points)** Let $Y_n = X_n - X_{n-1}$. Thus, $Y_n = 1$ indicates that the n th transition was to the right, $Y_n = 0$ indicates it was a self-transition, and $Y_n = -1$ indicates it was a transition to the left. Find $\lim_{n \rightarrow \infty} \mathbf{P}(Y_n = 1)$.

4. **(4 points)** Is the sequence Y_n a Markov chain? Justify your answer.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Final Exam | Fall 2010)

5. **(4 points)** Given that the n th transition was a transition to the right ($Y_n = 1$), find the probability that the previous state was state 1. (You can assume that n is large.)

6. **(4 points)** Suppose that $X_0 = 1$. Let T be defined as the first *positive time* at which the state is again equal to 1. Show how to find $\mathbf{E}[T]$. (It is enough to write down whatever equation(s) needs to be solved; you do not have to actually solve it/them or to produce a numerical answer.)

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Final Exam | Fall 2010)

7. **(4 points)** Does the sequence X_1, X_2, X_3, \dots converge in probability? If yes, to what? If not, just say “no” without explanation.

8. **(4 points)** Let $Z_n = \max\{X_1, \dots, X_n\}$. Does the sequence Z_1, Z_2, Z_3, \dots converge in probability? If yes, to what? If not, just say “no” without explanation.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Final Exam | Fall 2010)

Problem 2. (68 points) Alice shows up at an Athena* cluster at time zero and spends her time exclusively in typing emails. The times that her emails are sent are a Poisson process with rate λ_A per hour.

1. **(3 points)** What is the probability that Alice sent exactly three emails during the time interval $[1, 2]$?
 2. Let Y_1 and Y_2 be the times at which Alice's first and second emails were sent.
 - (a) **(3 points)** Find $\mathbf{E}[Y_2 \mid Y_1]$.

*Athena is MIT's UNIX-based computing environment. OCW does not provide access to it.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Final Exam | Fall 2010)

(b) **(3 points)** Find the PDF of Y_1^2 .

(c) **(3 points)** Find the joint PDF of Y_1 and Y_2 .

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Final Exam | Fall 2010)

3. You show up at time 1 and you are told that Alice has sent exactly one email so far. (Only give answers here, no need to justify them.)

(a) **(3 points)** What is the conditional expectation of Y_2 given this information?

(b) **(3 points)** What is the conditional expectation of Y_1 given this information?

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Final Exam | Fall 2010)

4. Bob just finished exercising (without email access) and sits next to Alice at time 1. He starts typing emails at time 1, and fires them according to an independent Poisson process with rate λ_B .
- (a) **(5 points)** What is the PMF of the total number of emails sent by the two of them together during the interval $[0, 2]$?

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Final Exam | Fall 2010)

- (b) **(5 points)** What is the expected value of the total typing time associated with the email that Alice is typing at the time that Bob shows up? (Here, “total typing time” includes the time that Alice spent on that email both before and after Bob’s arrival.)

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Final Exam | Fall 2010)

- (c) **(5 points)** What is the expected value of the time until each one of them has sent at least one email? (Note that we count time starting from time 0, and we take into account any emails possibly sent out by Alice during the interval $[0, 1]$.)

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Final Exam | Fall 2010)

- (d) **(5 points)** Given that a total of 10 emails were sent during the interval $[0, 2]$, what is the probability that exactly 4 of them were sent by Alice?

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Final Exam | Fall 2010)

5. **(5 points)** Suppose that $\lambda_A = 4$. Use Chebyshev's inequality to find an upper bound on the probability that Alice sent at least 5 emails during the time interval $[0, 1]$. Does the Markov inequality provide a better bound?
6. **(5 points)** You do not know λ_A but you watch Alice for an hour and see that she sent exactly 5 emails. Derive the maximum likelihood estimate of λ_A based on this information.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Final Exam | Fall 2010)

7. **(5 points)** We have reasons to believe that λ_A is a large number. Let N be the number of emails sent during the interval $[0, 1]$. Justify why the CLT can be applied to N , and give a precise statement of the CLT in this case.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Final Exam | Fall 2010)

8. **(5 points)** Under the same assumption as in last part, that λ_A is large, you can now pretend that N is a normal random variable. Suppose that you observe the value of N . Give an (approximately) 95% confidence interval for λ_A . State precisely what approximations you are making.
Possibly useful facts: The cumulative normal distribution satisfies $\Phi(1.645) = 0.95$ and $\Phi(1.96) = 0.975$.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Final Exam | Fall 2010)

9. You are now told that λ_A is actually the realized value of an exponential random variable Λ , with parameter 2:

$$f_{\Lambda}(\lambda) = 2e^{-2\lambda}, \quad \lambda \geq 0.$$

- (a) **(5 points)** Find $\mathbf{E}[N^2]$.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Final Exam | Fall 2010)

(b) **(5 points)** Find the linear least squares estimator of Λ given N .

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

4. (4 points) Is the sequence Y_n a Markov chain? Justify your answer.

Solution: No. Assume the Markov process is in steady state. To satisfy the Markov property,

$$\mathbf{P}(Y_n = 1 \mid Y_{n-1} = 1, Y_{n-2} = 1) = \mathbf{P}(Y_n = 1 \mid Y_{n-1} = 1).$$

For large n ,

$$\mathbf{P}(Y_n = 1 \mid Y_{n-1} = 1, Y_{n-2} = 1) = 0,$$

since it is not possible to move upwards 3 times in a row. However in steady state,

$$\begin{aligned}\mathbf{P}(Y_n = 1 \mid Y_{n-1} = 1) &= \frac{\mathbf{P}(\{Y_n = 1\} \cap \{Y_{n-1} = 1\})}{\mathbf{P}(Y_{n-1} = 1)} \\ &= \frac{\pi_1 p_{12} p_{23}}{\pi_1 p_{12} + \pi_2 p_{23}} \\ &\neq 0.\end{aligned}$$

Therefore, the sequence Y_n is not a Markov chain.

5. (4 points) Given that the n th transition was a transition to the right ($Y_n = 1$), find the probability that the previous state was state 1. (You can assume that n is large.)

Solution: Using Bayes' Rule,

$$\begin{aligned}\mathbf{P}(X_{n-1} = 1 \mid Y_n = 1) &= \frac{\mathbf{P}(X_{n-1} = 1)\mathbf{P}(Y_n = 1 \mid X_{n-1} = 1)}{\sum_{i=1}^3 \mathbf{P}(X_{n-1} = i)\mathbf{P}(Y_n = 1 \mid X_{n-1} = i)} \\ &= \frac{\pi_1 p_{12}}{\pi_1 p_{12} + \pi_2 p_{23}} \\ &= 2/5.\end{aligned}$$

6. (4 points) Suppose that $X_0 = 1$. Let T be defined as the first *positive time* at which the state is again equal to 1. Show how to find $\mathbf{E}[T]$. (It is enough to write down whatever equation(s) needs to be solved; you do not have to actually solve it/them or to produce a numerical answer.)

Solution: In order to find the mean recurrence time of state 1, the mean first passage times to state 1 are first calculated by solving the following system of equations:

$$\begin{aligned}t_2 &= 1 + p_{22}t_2 + p_{23}t_3 \\ t_3 &= 1 + p_{32}t_2 + p_{33}t_3.\end{aligned}$$

The mean recurrence time of state 1 is then $t_1^* = 1 + p_{12}t_2$.

Solving the system of equations yields $t_2 = 20$ and $t_3 = 30$ and $t_1^* = 9$.

7. (4 points) Does the sequence X_1, X_2, X_3, \dots converge in probability? If yes, to what? If not, just say "no" without explanation.

Solution: No.

8. (4 points) Let $Z_n = \max\{X_1, \dots, X_n\}$. Does the sequence Z_1, Z_2, Z_3, \dots converge in probability? If yes, to what? If not, just say “no” without explanation.

Solution: Yes. The sequence converges to 3 in probability.

For the original markov chain, states $\{1, 2, 3\}$ form one single recurrent class. Therefore, the Markov process will eventually visit each state with probability 1. In this case, the sequence Z_n will, with probability 1, converge to 3 once X_n visits 3 for the first time.

Problem 2. (68 points) Alice shows up at an Athena* cluster at time zero and spends her time exclusively in typing emails. The times that her emails are sent are a Poisson process with rate λ_A per hour.

1. (3 points) What is the probability that Alice sent exactly three emails during the time interval $[1, 2]$?

Solution: The number of emails Alice sends in the interval $[1, 2]$ is a Poisson random variable with parameter λ_A . So we have:

$$\mathbf{P}(3, 1) = \frac{\lambda_A^3 e^{-\lambda_A}}{3!}.$$

2. Let Y_1 and Y_2 be the times at which Alice’s first and second emails were sent.

- (a) (3 points) Find $\mathbf{E}[Y_2 | Y_1]$.

Solution: Define T_2 as the second inter-arrival time in Alice’s Poisson process. Then:

$$Y_2 = Y_1 + T_2$$

$$\mathbf{E}[Y_2 | Y_1] = \mathbf{E}[Y_1 + T_2 | Y_1] = Y_1 + \mathbf{E}[T_2] = Y_1 + 1/\lambda_A.$$

- (b) (3 points) Find the PDF of Y_1^2 .

Solution: Let $Z = Y_1^2$. Then we first find the CDF of Z and differentiate to find the PDF of Z :

$$F_Z(z) = \mathbf{P}(Y_1^2 \leq z) = \mathbf{P}(-\sqrt{z} \leq Y_1 \leq \sqrt{z}) = \begin{cases} 1 - e^{-\lambda_A \sqrt{z}} & z \geq 0 \\ 0 & z < 0. \end{cases}$$

$$\begin{aligned} f_Z(z) &= \frac{dF_Z(z)}{dz} = \lambda_A e^{-\lambda_A \sqrt{z}} \left(\frac{1}{2} z^{-1/2} \right) && (z \geq 0) \\ f_Z(z) &= \begin{cases} \frac{\lambda_A}{2\sqrt{z}} e^{-\lambda_A \sqrt{z}} & z \geq 0 \\ 0 & z < 0. \end{cases} \end{aligned}$$

- (c) (3 points) Find the joint PDF of Y_1 and Y_2 .

Solution:

$$\begin{aligned} f_{Y_1, Y_2}(y_1, y_2) &= f_{Y_1}(y_1) f_{Y_2|Y_1}(y_2|y_1) \\ &= f_{Y_1}(y_1) f_{T_2}(y_2 - y_1) \\ &= \lambda_A e^{-\lambda_A y_1} \lambda_A e^{-\lambda_A (y_2 - y_1)} && y_2 \geq y_1 \geq 0 \\ &= \begin{cases} \lambda_A^2 e^{-\lambda_A y_2} & y_2 \geq y_1 \geq 0 \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

3. You show up at time 1 and you are told that Alice has sent exactly one email so far. (Only give answers here, no need to justify them.)

- (a) **(3 points)** What is the conditional expectation of Y_2 given this information?

Solution: Let A be the event {exactly one arrival in the interval $[0,1]$ }. Looking forward from time $t = 1$, the time until the next arrival is simply an exponential random variable (T). So,

$$\mathbf{E}[Y_2 | A] = 1 + \mathbf{E}[T] = 1 + 1/\lambda_A.$$

- (b) **(3 points)** What is the conditional expectation of Y_1 given this information?

Solution: Given A , the times in this interval are equally likely for the arrival Y_1 . Thus,

$$\mathbf{E}[Y_1 | A] = 1/2.$$

4. Bob just finished exercising (without email access) and sits next to Alice at time 1. He starts typing emails at time 1, and fires them according to an independent Poisson process with rate λ_B .

- (a) **(5 points)** What is the PMF of the total number of emails sent by the two of them together during the interval $[0, 2]$?

Solution: Let K be the total number of emails sent in $[0, 2]$. Let K_1 be the total number of emails sent in $[0, 1]$, and let K_2 be the total number of emails sent in $[1, 2]$. Then $K = K_1 + K_2$ where K_1 is a Poisson random variable with parameter λ_A and K_2 is a Poisson random variable with parameter $\lambda_A + \lambda_B$ (since the emails sent by both Alice and Bob after time $t = 1$ arrive according to the merged Poisson process of Alice's emails and Bob's emails). Since K is the sum of independent Poisson random variables, K is a Poisson random variable with parameter $2\lambda_A + \lambda_B$. So K has the distribution:

$$p_K(k) = \frac{(2\lambda_A + \lambda_B)^k e^{-(2\lambda_A + \lambda_B)}}{k!} \quad k = 0, 1, \dots$$

- (b) **(5 points)** What is the expected value of the total typing time associated with the email that Alice is typing at the time that Bob shows up? (Here, “total typing time” includes the time that Alice spent on that email both before and after Bob’s arrival.)

Solution: The total typing time Q associated with the email that Alice is typing at the time Bob shows up is the sum of S_0 , the length of time between Alice's last email or time 0 (whichever is later) and time 1, and T_1 , the length of time from 1 to the time at which Alice sends her current email. T_1 is exponential with parameter λ_A . and $S_0 = \min\{T_0, 1\}$, where T_0 is exponential with parameter λ_A .

Then,

$$Q = S_0 + T_1 = \min\{T_0, 1\} + T_1$$

and

$$\mathbf{E}[Q] = \mathbf{E}[S_0] + \mathbf{E}[T_1].$$

We have: $\mathbf{E}[T_1] = 1/\lambda_A$.

We can find $\mathbf{E}[S_0]$ via the law of total expectations:

$$\begin{aligned}
 \mathbf{E}[S_0] = \mathbf{E}[\min\{T_0, 1\}] &= \mathbf{P}(T_0 \leq 1)\mathbf{E}[T_0 | T_0 \leq 1] + \mathbf{P}(T_0 > 1)\mathbf{E}[1|T_0 > 1] \\
 &= (1 - e^{-\lambda_A}) \int_0^1 t f_{T|T_0 \leq 1}(t) dt + e^{-\lambda_A} \\
 &= (1 - e^{-\lambda_A}) \int_0^1 t \frac{\lambda_A e^{-\lambda_A t}}{(1 - e^{-\lambda_A})} dt + e^{-\lambda_A} \\
 &= \int_0^1 t \lambda_A e^{-\lambda_A t} dt + e^{-\lambda_A} \\
 &= \frac{1}{\lambda_A} \int_0^1 t \lambda_A^2 e^{-\lambda_A t} dt + e^{-\lambda_A} \\
 &= \frac{1}{\lambda_A} (1 - e^{-\lambda_A} - \lambda_A e^{-\lambda_A}) + e^{-\lambda_A} \\
 &= \frac{1}{\lambda_A} (1 - e^{-\lambda_A})
 \end{aligned}$$

where the above integral is evaluated by manipulating the integrand into an Erlang order 2 PDF and equating the integral of this PDF from 0 to 1 to the probability that there are 2 or more arrivals in the first hour (i.e. $\mathbf{P}(Y_2 < 1) = 1 - \mathbf{P}(0, 1) - \mathbf{P}(1, 1)$). Alternatively, one can integrate by parts and arrive at the same result.

Combining the above expectations:

$$\mathbf{E}[Q] = \mathbf{E}[S_0] + \mathbf{E}[T_1] = \frac{1}{\lambda_A} (1 - e^{-\lambda_A}) + \frac{1}{\lambda_A} = \frac{1}{\lambda_A} (2 - e^{-\lambda_A}).$$

- (c) **(5 points)** What is the expected value of the time until each one of them has sent at least one email? (Note that we count time starting from time 0, and we take into account any emails possibly sent out by Alice during the interval $[0, 1]$.)

Solution: Define U as the time from $t = 0$ until each person has sent at least one email. Define V as the remaining time from when Bob arrives (time 1) until each person has sent at least one email (so $V = U - 1$).

Define S as the time until Bob sends his first email after time 1.

Define the event $A = \{\text{Alice sends one or more emails in the time interval } [0, 1]\} = \{Y_1 \leq 1\}$, where Y_1 is the time Alice sends her first email.

Define the event $B = \{\text{After time 1, Bob sends his next email before Alice does}\}$, which is equivalent to the event where the next arrival in the merged process from Alice and Bob's original processes (starting from time 1) comes from Bob's process.

We have:

$$\mathbf{P}(A) = \mathbf{P}(Y_1 \leq 1) = 1 - e^{-\lambda_A}$$

$$\mathbf{P}(B) = \frac{\lambda_B}{\lambda_A + \lambda_B}.$$

Then,

$$\begin{aligned}
 \mathbf{E}[U] &= \mathbf{P}(A)\mathbf{E}[U | A] + \mathbf{P}(A^c)\mathbf{E}[U | A^c] \\
 &= (1 - e^{-\lambda_A})(1 + \mathbf{E}[V | A]) + e^{-\lambda_A}(1 + \mathbf{E}[V | A^c]) \\
 &= (1 - e^{-\lambda_A})(1 + \mathbf{E}[V | A]) + e^{-\lambda_A}(1 + \mathbf{P}(B | A^c)\mathbf{E}[V | B \cap A^c] + \mathbf{P}(B^c | A^c)\mathbf{E}[V | B^c \cap A^c]) \\
 &= (1 - e^{-\lambda_A})(1 + \mathbf{E}[V | A]) + e^{-\lambda_A}(1 + \mathbf{P}(B)\mathbf{E}[V | B \cap A^c] + \mathbf{P}(B^c)\mathbf{E}[V | B^c \cap A^c]) \\
 &= (1 - e^{-\lambda_A})(1 + \mathbf{E}[V | A]) + e^{-\lambda_A} \left(1 + \frac{\lambda_B}{\lambda_A + \lambda_B}\mathbf{E}[V | B \cap A^c] + \frac{\lambda_A}{\lambda_A + \lambda_B}\mathbf{E}[V | B^c \cap A^c] \right).
 \end{aligned}$$

Note that $\mathbf{E}[V | B^c \cap A^c]$ is the expected value of the time until each of them sends one email after time 1 (since, given A^c , Alice did not send any in the interval $[0, 1]$) and given Alice sends an email before Bob. Then this is the expected time until an arrival in the merged process followed by the expected time until an arrival in Bob's process. So, $\mathbf{E}[V | B^c \cap A^c] = \frac{1}{\lambda_A + \lambda_B} + \frac{1}{\lambda_B}$.

Similarly, $\mathbf{E}[V | B \cap A^c]$ is the time until each sends an email after time 1, given Bob sends an email before Alice. So $\mathbf{E}[V | B \cap A^c] = \frac{1}{\lambda_A + \lambda_B} + \frac{1}{\lambda_A}$.

Also, $\mathbf{E}[V | A]$ is the expected time it takes for Bob to send his first email after time 1 (since, given A , Alice already sent an email in the interval $[0, 1]$). So $\mathbf{E}[V | A] = \mathbf{E}[S] = 1/\lambda_B$. Combining all of this with the above, we have:

$$\begin{aligned}
 \mathbf{E}[U] &= (1 - e^{-\lambda_A})(1 + 1/\lambda_B) \\
 &\quad + e^{-\lambda_A} \left(1 + \frac{\lambda_B}{\lambda_A + \lambda_B} \left(\frac{1}{\lambda_A + \lambda_B} + \frac{1}{\lambda_B} \right) + \frac{\lambda_A}{\lambda_A + \lambda_B} \left(\frac{1}{\lambda_A + \lambda_B} + \frac{1}{\lambda_A} \right) \right).
 \end{aligned}$$

- (d) **(5 points)** Given that a total of 10 emails were sent during the interval $[0, 2]$, what is the probability that exactly 4 of them were sent by Alice?

Solution:

$$\begin{aligned}
 \mathbf{P}(\text{Alice sent 4 in } [0, 2] | \text{total 10 sent in } [0, 2]) &= \frac{\mathbf{P}(\text{Alice sent 4 in } [0, 2] \cap \text{total 10 sent in } [0, 2])}{\mathbf{P}(\text{total 10 sent in } [0, 2])} \\
 &= \frac{\mathbf{P}(\text{Alice sent 4 in } [0, 2] \cap \text{Bob sent 6 in } [0, 2])}{\mathbf{P}(\text{total 10 sent in } [0, 2])} \\
 &= \frac{\left(\frac{(2\lambda_A)^4 e^{-2\lambda_A}}{4!} \right) \left(\frac{(\lambda_B)^6 e^{-\lambda_B}}{6!} \right)}{\frac{(2\lambda_A + \lambda_B)^{10} e^{-2\lambda_A - \lambda_B}}{10!}} \\
 &= \binom{10}{4} \left(\frac{2\lambda_A}{2\lambda_A + \lambda_B} \right)^4 \left(\frac{\lambda_B}{2\lambda_A + \lambda_B} \right)^6.
 \end{aligned}$$

As the form of the solution suggests, the problem can be solved alternatively by computing the probability of a single email being sent by Alice, given it was sent in the interval $[0, 2]$. This can be found by viewing the number of emails sent by Alice in $[0, 2]$ as the number of arrivals arising from a Poisson process with twice the rate ($2\lambda_A$) in an interval of half the duration (particularly, the interval $[1, 2]$), then merging this process with Bob's process. Then the probability that an email sent in the interval $[0, 2]$ was sent by Alice is the probability that an arrival in this new merged process came from the newly constructed $2\lambda_A$ rate process:

$$p = \frac{2\lambda_A}{2\lambda_A + \lambda_B}.$$

Then, out of 10 emails, the probability that 4 came from Alice is simply a binomial probability with 4 successes in 10 trials, which agrees with the solution above.

5. **(5 points)** Suppose that $\lambda_A = 4$. Use Chebyshev's inequality to find an upper bound on the probability that Alice sent at least 5 emails during the time interval $[0, 1]$. Does the Markov inequality provide a better bound?

Solution:

Let N be the number of emails Alice sent in the interval $[0, 1]$. Since N is a Poisson random variable with parameter λ_A ,

$$\mathbf{E}[N] = \text{var}(N) = \lambda_A = 4.$$

To apply the Chebyshev inequality, we recognize:

$$\mathbf{P}(N \geq 5) = \mathbf{P}(N - 4 \geq 1) \leq \mathbf{P}(|N - 4| \geq 1) \leq \frac{\text{var}(N)}{1^2} = 4.$$

In this case, the upper-bound of 4 found by application of the Chebyshev inequality is uninformative, as we already knew $\mathbf{P}(N \geq 5) \leq 1$.

To find a better bound on this probability, use the Markov inequality, which gives:

$$\mathbf{P}(N \geq 5) \leq \frac{\mathbf{E}[N]}{5} = \frac{4}{5}.$$

6. **(5 points)** You do not know λ_A but you watch Alice for an hour and see that she sent exactly 5 emails. Derive the maximum likelihood estimate of λ_A based on this information.

Solution:

$$\begin{aligned}\hat{\lambda}_A &= \arg \max_{\lambda} \log(p_N(5; \lambda)) \\ &= \arg \max_{\lambda} \log\left(\frac{\lambda^5 e^{-\lambda}}{5!}\right) \\ &= \arg \max_{\lambda} -\log(5!) + 5\log(\lambda) - \lambda.\end{aligned}$$

Setting the first derivative to zero

$$\frac{5}{\lambda} - 1 = 0$$

$$\hat{\lambda}_A = 5.$$

7. **(5 points)** We have reasons to believe that λ_A is a large number. Let N be the number of emails sent during the interval $[0, 1]$. Justify why the CLT can be applied to N , and give a precise statement of the CLT in this case.

Solution: With λ_A large, we assume $\lambda_A \gg 1$. For simplicity, assume λ_A is an integer. We can divide the interval $[0, 1]$ into λ_A disjoint intervals, each with duration $1/\lambda_A$, so that these intervals span the entire interval from $[0, 1]$. Let N_i be the number of arrivals in the i th such interval, so that the N_i 's are independent, identically distributed Poisson random variables with parameter 1. Since N is defined as the number of arrivals in the interval $[0, 1]$, then $N = N_1 + \dots + N_{\lambda_A}$. Since $\lambda_A \gg 1$, then N is the sum of a large number of independent and identically distributed random variables, where the distribution of N_i does not change as the number of terms in the sum increases. Hence, N is approximately normal with mean λ_A and variance λ_A .

If λ_A is not an integer, the same argument holds, except that instead of having λ_A intervals, we have an integer number of intervals equal to the integer part of λ_A ($\bar{\lambda}_A = \text{floor}(\lambda_A)$) of length $1/\lambda_A$ and an extra interval of a shorter length $(\lambda_A - \bar{\lambda}_A)/\lambda_A$.

Now, N is a sum of λ_A independent, identically distributed Poisson random variables with parameter 1 added to another Poisson random variable (also independent of all the other Poisson random variables) with parameter $(\lambda_A - \bar{\lambda}_A)$. In this case, N would need a small correction to apply the central limit theorem as we are familiar with it; however, it turns out that even without this correction, adding the extra Poisson random variable does not preclude the distribution of N from being approximately normal, for large λ_A , and the central limit theorem still applies.

To arrive at a precise statement of the CLT, we must “standardize” N by subtracting its mean then dividing by its standard deviation. After having done so, the CDF of the standardized version of N should converge to the standard normal CDF as the number of terms in the sum approaches infinity (as $\lambda_A \rightarrow \infty$).

Therefore, the precise statement of the CLT when applied to N is:

$$\lim_{\lambda_A \rightarrow \infty} \mathbf{P} \left(\frac{N - \lambda_A}{\sqrt{\lambda_A}} \leq z \right) = \Phi(z)$$

where $\Phi(z)$ is the standard normal CDF.

8. **(5 points)** Under the same assumption as in last part, that λ_A is large, you can now pretend that N is a normal random variable. Suppose that you observe the value of N . Give an (approximately) 95% confidence interval for λ_A . State precisely what approximations you are making.

Possibly useful facts: The cumulative normal distribution satisfies $\Phi(1.645) = 0.95$ and $\Phi(1.96) = 0.975$.

Solution: We begin by estimating λ_A with its ML estimator $\hat{\lambda}_A = N$, where $\mathbf{E}[N] = \lambda_A$. With λ_A large, the CLT applies, and we can assume N has an approximately normal distribution. Since $\text{var}(N) = \lambda_A$, we can also approximate the variance of N with ML estimator for λ_A , so $\text{var}(N) \approx N$, and $\sigma_N \approx \sqrt{N}$.

To find the 95% confidence interval, we find β such that:

$$\begin{aligned} 0.95 &= \mathbf{P}(|N - \lambda_A| \leq \beta) \\ &= \mathbf{P}\left(\frac{|N - \lambda_A|}{\sqrt{N}} \leq \frac{\beta}{\sqrt{N}}\right) \\ &\approx 2\Phi\left(\frac{\beta}{\sqrt{N}}\right). \end{aligned}$$

So, we find:

$$\beta \approx \sqrt{N}\Phi^{-1}(0.975) = 1.96\sqrt{N}.$$

Thus, we can write:

$$\mathbf{P}(N - 1.96\sqrt{N} \leq \lambda_A \leq N + 1.96\sqrt{N}) \approx 0.95.$$

So, the approximate 95% confidence interval is: $[N - 1.96\sqrt{N}, N + 1.96\sqrt{N}]$.

9. You are now told that λ_A is actually the realized value of an exponential random variable Λ , with parameter 2:

$$f_\Lambda(\lambda) = 2e^{-2\lambda}, \quad \lambda \geq 0.$$

- (a) (5 points) Find $\mathbf{E}[N^2]$.

Solution:

$$\begin{aligned} \mathbf{E}[N^2] &= \mathbf{E}[\mathbf{E}[N^2 | \Lambda]] = \mathbf{E}[\text{var}(N | \Lambda) + (\mathbf{E}[N | \Lambda])^2] \\ &= \mathbf{E}[\Lambda + \Lambda^2] \\ &= \mathbf{E}[\Lambda] + \text{var}(\Lambda) + (\mathbf{E}[\Lambda])^2 \\ &= \frac{1}{2} + \frac{2}{2^2} \\ &= 1. \end{aligned}$$

- (b) (5 points) Find the linear least squares estimator of Λ given N .

Solution:

$$\hat{\Lambda}_{\text{LLMS}} = \mathbf{E}[\Lambda] + \frac{\text{cov}(N, \Lambda)}{\text{var}(N)}(N - \mathbf{E}[N]).$$

Solving for the above quantities:

$$\mathbf{E}[\Lambda] = \frac{1}{2}$$

$$\mathbf{E}[N] = \mathbf{E}[\mathbf{E}[N | \Lambda]] = \mathbf{E}[\Lambda] = \frac{1}{2}.$$

$$\text{var}(N) = \mathbf{E}[N^2] - (\mathbf{E}[N])^2 = 1 - \frac{1}{2^2} = \frac{3}{4}.$$

$$\text{cov}(N, \Lambda) = \mathbf{E}[N\Lambda] - \mathbf{E}[N]\mathbf{E}[\Lambda] = \mathbf{E}[\mathbf{E}[N\Lambda | \Lambda]] - (\mathbf{E}[\Lambda])^2 = \mathbf{E}[\Lambda^2] - (\mathbf{E}[\Lambda])^2 = \text{var}(\Lambda) = \frac{1}{4}.$$

Substituting these into the equation above:

$$\begin{aligned}\hat{\Lambda}_{\text{LLMS}} &= \mathbf{E}[\Lambda] + \frac{\text{cov}(N, \Lambda)}{\text{var}(N)}(N - \mathbf{E}[N]) \\ &= \frac{1}{2} + \frac{1/4}{3/4} \left(N - \frac{1}{2} \right) \\ &= \frac{1}{3} (N + 1).\end{aligned}$$

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Problem 2. (20 points)

A pair of jointly continuous random variables, X and Y , have a joint probability density function given by

$$f_{X,Y}(x,y) = \begin{cases} c, & \text{in the shaded region of Fig. 1} \\ 0, & \text{elsewhere.} \end{cases}$$

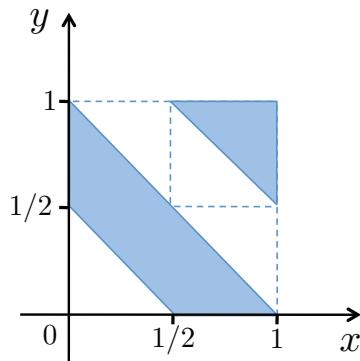
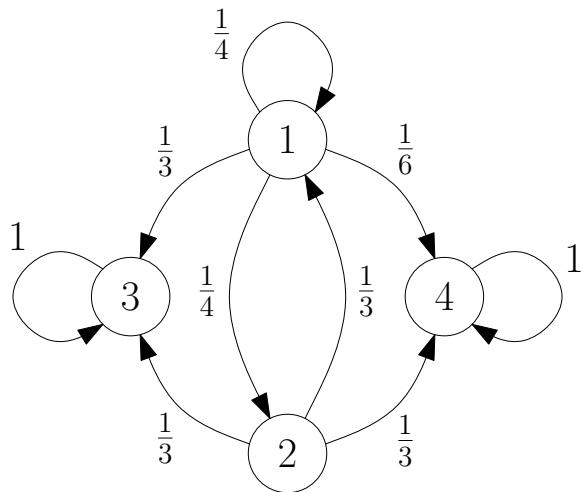


Figure 1: The shaded region is the domain in which $f_{X,Y}(x,y) = c$.

- (a) (5 points) Find c .
- (b) (5 points) Find the marginal PDFs of X and Y , i.e., $f_X(x)$ and $f_Y(y)$.
- (c) (5 points) Find $\mathbf{E}[X \mid Y = 1/4]$ and $\text{Var}[X \mid Y = 1/4]$, that is, the conditional mean and conditional variance of X given that $Y = 1/4$.
- (d) (5 points) Find the conditional PDF for X given that $Y = 3/4$, i.e., $f_{X|Y}(x \mid 3/4)$.

Problem 3. (25 points)

Consider a Markov chain X_n whose one-step transition probabilities are shown in the figure.



- (a) (5 points) What are the recurrent states?

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Final Exam | Fall 2009)

- (b) (5 points) Find $\mathbf{P}(X_2 = 4 \mid X_0 = 2)$.
- (c) (5 points) Suppose that you are given the values of $r_{ij}(n) = \mathbf{P}(X_n = j \mid X_0 = i)$. Give a formula for $r_{11}(n+1)$ in terms of the $r_{ij}(n)$.
- (d) (5 points) Find the steady-state probabilities $\pi_j = \lim_{n \rightarrow \infty} \mathbf{P}(X_n = j \mid X_0 = i)$, or explain why they do not exist.
- (e) (5 points) What is the probability of eventually visiting state 4, given that the initial state is $X_0 = 1$?

Problem 4. (30 points)

Al, Bonnie, and Clyde run laps around a track, with the duration of each lap (in hours) being exponentially distributed with parameters $\lambda_A = 21$, $\lambda_B = 23$, and $\lambda_C = 24$, respectively. Assume that all lap durations are independent. At the completion of each lap, a runner drinks either one or two cups of water, with probabilities $1/3$ and $2/3$, respectively, independent of everything else, including how much water was consumed after previous laps. (The time spent drinking is negligible, assumed zero.)

- (a) (5 points) Write down the PMF of the total number of completed laps over the first hour.
- (b) (5 points) What is the expected number of cups of water to be consumed by the three runners, in total, over the first hour.
- (c) (5 points) Al has amazing endurance and completed 72 laps. Find a good approximation for the probability that he drank at least 130 cups. (You do not have to use 1/2-corrections.)
- (d) (5 points) What is the probability that Al finishes his first lap before any of the others?
- (e) (5 points) Suppose that the runners have been running for a very long time when you arrive at the track. What is the distribution of the duration of Al's current lap? (This includes the duration of that lap both before and after the time of your arrival.)
- (f) (5 points) Suppose that the runners have been running for $1/4$ hours. What is the distribution of the time Al spends on his second lap, given that he is on his second lap?

Problem 5. (25 points)

A pulse of light has energy X that is a second-order Erlang random variable with parameter λ , i.e., its PDF is

$$f_X(x) = \begin{cases} \lambda^2 x e^{-\lambda x}, & \text{for } x \geq 0, \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

This pulse illuminates an ideal photon-counting detector whose output N is a Poisson-distributed random variable with mean x when $X = x$, i.e., its conditional PMF is

$$p_{N|X}(n \mid x) = \begin{cases} \frac{x^n e^{-x}}{n!}, & \text{for } n = 0, 1, 2, \dots, \\ 0, & \text{otherwise.} \end{cases}$$

- (a) (5 points) Find $\mathbf{E}[N]$ and $\text{Var}[N]$, the unconditional mean and variance of N .
- (b) (5 points) Find $p_N(n)$, the unconditional PMF of N .
- (c) (5 points) Find $\hat{X}_{\text{lin}}(N)$, the linear least-squares estimator of X based on an observation of N .
- (d) (5 points) Find $\hat{X}_{\text{MAP}}(N)$, the MAP estimator of X based on an observation of N .

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Final Exam | Fall 2009)

- (e) (5 points) Instead of the prior distribution in Eq. (1), we are now told that

$$\mathbf{P}(X = 2) = 3^3/35, \quad \mathbf{P}(X = 3) = 2^3/35.$$

Given the observation $N = 3$, and in order to minimize the probability of error, which one of the two hypotheses $X = 2$ and $X = 3$ should be chosen?

Useful integral and facts:

$$\int_0^\infty y^k e^{-\alpha y} dy = \frac{k!}{\alpha^{k+1}}, \quad \text{for } \alpha > 0 \text{ and } k = 0, 1, 2, \dots \text{ (recall that } 0!=1\text{)}$$

The second-order Erlang random variable satisfies:

$$\mathbf{E}[X] = 2/\lambda, \quad \text{Var}(X) = 2/\lambda^2.$$

Each question is repeated in the following pages. Please write your answer on the appropriate page.

Problem 2. (20 points)

A pair of jointly continuous random variables, X and Y , have a joint probability density function given by

$$f_{X,Y}(x,y) = \begin{cases} c, & \text{in the shaded region of Fig. 1} \\ 0, & \text{elsewhere.} \end{cases}$$

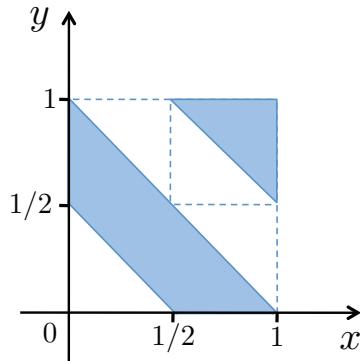


Figure 2: The shaded region is the domain in which $f_{X,Y}(x,y) = c$.

- (a) (5 points) Find c .

- (b) (5 points) Find the marginal PDFs of X and Y , i.e., $f_X(x)$ and $f_Y(y)$.

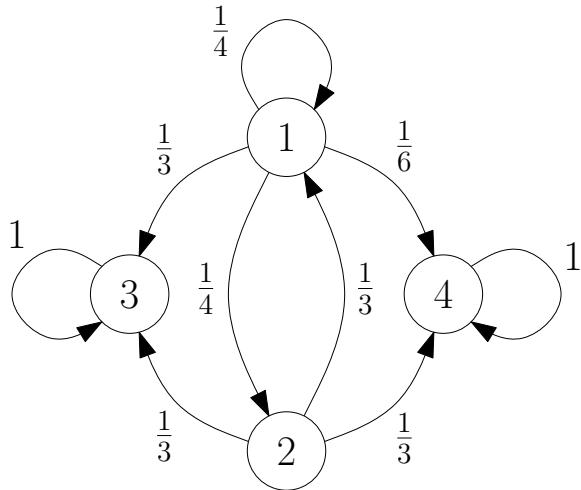
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Final Exam | Fall 2009)

(c) (5 points) Find $\mathbf{E}[X \mid Y = 1/4]$ and $\text{Var}[X \mid Y = 1/4]$, that is, the conditional mean and conditional variance of X given that $Y = 1/4$.

(d) (5 points) Find the conditional PDF for X given that $Y = 3/4$, i.e., $f_{X|Y}(x \mid 3/4)$.

Problem 3. (25 points)

Consider a Markov chain X_n whose one-step transition probabilities are shown in the figure.



(a) (5 points) What are the recurrent states?

(b) (5 points) Find $\mathbf{P}(X_2 = 4 \mid X_0 = 2)$.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Final Exam | Fall 2009)

(c) (5 points) Suppose that you are given the values of $r_{ij}(n) = \mathbf{P}(X_n = j \mid X_0 = i)$. Give a formula for $r_{11}(n+1)$ in terms of the $r_{ij}(n)$.

(d) (5 points) Find the steady-state probabilities $\pi_j = \lim_{n \rightarrow \infty} \mathbf{P}(X_n = j \mid X_0 = i)$, or explain why they do not exist.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Final Exam | Fall 2009)

- (e) (5 points) What is the probability of eventually visiting state 4, given that the initial state is $X_0 = 1$?

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Final Exam | Fall 2009)

Problem 4. (30 points)

Al, Bonnie, and Clyde run laps around a track, with the duration of each lap (in hours) being exponentially distributed with parameters $\lambda_A = 21$, $\lambda_B = 23$, and $\lambda_C = 24$, respectively. Assume that all lap durations are independent. At the completion of each lap, a runner drinks either one or two cups of water, with probabilities $1/3$ and $2/3$, respectively, independent of everything else, including how much water was consumed after previous laps. (The time spent drinking is negligible, assumed zero.)

(a) (5 points) Write down the PMF of the total number of completed laps over the first hour.

(b) (5 points) What is the expected number of cups of water to be consumed by the three runners, in total, over the first hour.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Final Exam | Fall 2009)

(c) (5 points) Al has amazing endurance and completed 72 laps. Find a good approximation for the probability that he drank at least 130 cups. (You do not have to use 1/2-corrections.)

(d) (5 points) What is the probability that Al finishes his first lap before any of the others?

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Final Exam | Fall 2009)

- (e) (5 points) Suppose that the runners have been running for a very long time when you arrive at the track. What is the distribution of the duration of Al's current lap? (This includes the duration of that lap both before and after the time of your arrival.)
- (f) (5 points) Suppose that the runners have been running for $1/4$ hours. What is the distribution of the time Al spends on his second lap, given that he is on his second lap?

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Final Exam | Fall 2009)

Problem 5. (25 points)

A pulse of light has energy X that is a second-order Erlang random variable with parameter λ , i.e., its PDF is

$$f_X(x) = \begin{cases} \lambda^2 x e^{-\lambda x}, & \text{for } x \geq 0, \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

This pulse illuminates an ideal photon-counting detector whose output N is a Poisson-distributed random variable with mean x when $X = x$, i.e., its conditional PMF is

$$p_{N|X}(n | x) = \begin{cases} \frac{x^n e^{-x}}{n!}, & \text{for } n = 0, 1, 2, \dots, \\ 0, & \text{otherwise.} \end{cases}$$

Useful integral and facts:

$$\int_0^\infty y^k e^{-\alpha y} dy = \frac{k!}{\alpha^{k+1}}, \quad \text{for } \alpha > 0 \text{ and } k = 0, 1, 2, \dots \text{ (recall that } 0!=1\text{)}$$

The second-order Erlang random variable satisfies:

$$\mathbf{E}[X] = 2/\lambda, \quad \text{Var}(X) = 2/\lambda^2.$$

- (a) (5 points) Find $\mathbf{E}[N]$ and $\text{Var}[N]$, the unconditional mean and variance of N

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Final Exam | Fall 2009)

(b) (5 points) Find $p_N(n)$, the unconditional PMF of N .

(c) (5 points) Find $\hat{X}_{\text{lin}}(N)$, the linear least-squares estimator of X based on an observation of N .

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Final Exam | Fall 2009)

(d) (5 points) Find $\hat{X}_{\text{MAP}}(N)$, the MAP estimator of X based on an observation of N .

(e) (5 points) Instead of the prior distribution in Eq. (1), we are now told that

$$\mathbf{P}(X = 2) = 3^3/35, \quad \mathbf{P}(X = 3) = 2^3/35.$$

Given the observation $N = 3$, and in order to minimize the probability of error, which one of the two hypotheses $X = 2$ and $X = 3$ should be chosen?

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

Problem 3. (25 points)

(a) (5 points)

The recurrent states are {3,4}.

(b) (5 points)

The 2-step transition probability from State 2 to State 4 can be found by enumerating all the possible sequences. They are {2 → 1 → 4} and {2 → 4 → 4}. Thus,

$$\mathbf{P}(X_2 = 4 \mid X_0 = 2) = \frac{1}{3} \cdot \frac{1}{6} + \frac{1}{3} \cdot 1 = \frac{7}{18}.$$

(c) (5 points)

Generally,

$$r_{11}(n+1) = \sum_{j=1}^4 p_{1j} r_{j1}(n).$$

Since states 3 and 4 are absorbing states, this expression simplifies to

$$r_{11}(n+1) = \frac{1}{4} r_{11}(n) + \frac{1}{4} r_{21}(n).$$

Alternatively,

$$\begin{aligned} r_{11}(n+1) &= \sum_{k=1}^4 r_{1k}(n) p_{k1} \\ &= r_{11}(n) \cdot \frac{1}{4} + r_{12}(n) \cdot \frac{1}{3}. \end{aligned}$$

(d) (5 points)

The steady-state probabilities do not exist since there is more than one recurrent class. The long-term state probabilities would depend on the initial state.

(e) (5 points)

To find the probability of being absorbed by state 4, we set up the absorption probabilities. Note that $a_4 = 1$ and $a_3 = 0$.

$$\begin{aligned} a_1 &= \frac{1}{4}a_1 + \frac{1}{4}a_2 + \frac{1}{3}a_3 + \frac{1}{6}a_4 \\ &= \frac{1}{4}a_1 + \frac{1}{4}a_2 + \frac{1}{6} \end{aligned}$$

$$\begin{aligned} a_2 &= \frac{1}{3}a_1 + \frac{1}{3}a_3 + \frac{1}{3}a_4 \\ &= \frac{1}{3}a_1 + \frac{1}{3} \end{aligned}$$

Solving these equations yields $a_1 = \frac{3}{8}$.

Problem 4. (30 points)

(a) (5 points)

Given the problem statement, we can treat Al, Bonnie, and Clyde's running as 3 independent Poisson processes, where the arrivals correspond to lap completions and the arrival rates indicate the number of laps completed per hour. Since the three processes are independent, we can merge them to create a new process that captures the lap completions of all three runners. This merged process will have arrival rate $\lambda_M = \lambda_A + \lambda_B + \lambda_C = 68$. The total number of completed laps, L , over the first hour is then described by a Poisson PMF with $\lambda_M = 68$ and $\tau = 1$:

$$p_L(\ell) = \begin{cases} \frac{68^\ell e^{-68}}{\ell!}, & \ell = 0, 1, 2, \dots, \\ 0, & \text{otherwise.} \end{cases}$$

(b) (5 points)

Let L be the total number of completed laps over the first hour, and let C_i be the number of cups of water consumed at the end of the i th lap. Then, the total number of cups of water consumed is

$$C = \sum_{i=1}^L C_i,$$

which is a sum of a random number of i.i.d. random variables. Thus, we can use the law of iterated expectations to find

$$\mathbf{E}[C] = \mathbf{E}[\mathbf{E}[C | L]] = \mathbf{E}[LC_i] = \mathbf{E}[L]\mathbf{E}[C_i] = (\lambda_M\tau) \cdot \left(1 \cdot \frac{1}{3} + 2 \cdot \frac{2}{3}\right) = 68 \cdot \frac{5}{3} = \frac{340}{3}.$$

(c) (5 points)

Let X be the number of laps (out of 72) after which Al drank 2 cups of water. Then, in order for him to drink at least 130 cups, we must have

$$1 \cdot (72 - X) + 2 \cdot X \geq 130,$$

which implies that we need

$$X \geq 58.$$

Now, let X_i be i.i.d. Bernoulli random variables that equal 1 if Al drank 2 cups of water following his i th lap and 0 if he drank 1 cup. Then

$$X = X_1 + X_2 + \cdots + X_{72}.$$

X is evidently a binomial random variable with $n = 72$ and $p = 2/3$, and the probability we are looking for is

$$\mathbf{P}(X \geq 58) = \sum_{k=58}^{72} \binom{72}{k} \left(\frac{2}{3}\right)^k \left(\frac{1}{3}\right)^{72-k}.$$

This expression is difficult to calculate, but since we're dealing with the sum of a relatively large number of i.i.d. random variables, we can invoke the Central Limit Theorem to approximate this probability using a normal distribution. In particular, we can approximate X as

a normal random variable with mean $np = 72 \cdot 2/3 = 48$ and variance $np(1-p) = 16$ and approximate the desired probability as

$$\mathbf{P}(X \geq 58) = 1 - \mathbf{P}(X < 58) \approx 1 - \Phi\left(\frac{58 - 48}{\sqrt{16}}\right) = 1 - \Phi(2.5) \approx 0.0062.$$

(d) (5 points)

The event that Al is the first to finish a lap is the same as the event that the first arrival in the merged process came from Al's process. This probability is

$$\frac{\lambda_A}{\lambda_A + \lambda_B + \lambda_C} = \frac{21}{68}.$$

(e) (5 points)

This is an instance of the random incidence paradox, so the duration of Al's current lap consists of the sum of the duration from the time of your arrival until Al's next lap completion and the duration from the time of your arrival back to the time of Al's previous lap completion. This is the sum of 2 independent exponential random variables with parameter $\lambda_A = 21$ (i.e. a second- order Erlang random variable):

$$f_T(t) = \begin{cases} 21^2 t e^{-21t}, & t \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

(f) (5 points)

As in the previous part, the duration of Al's second lap consists of the time remaining from $t = 1/4$ until he completes his second lap and the time elapsed since he began his second lap until $t = 1/4$. Let X be the time elapsed and Y be the time remaining. We can still model the time remaining Y as an exponential random variable. However, we can no longer do the same for the time elapsed X because we know X can be no larger than $1/4$, whereas the exponential random variable can be arbitrarily large.

To find the PDF of X , let's first consider its CDF.

$$\begin{aligned} \mathbf{P}(X \leq x) &= \mathbf{P}(\text{The 1 arrival occurred less than } x \text{ hours ago from time } 1/4) \\ &= \frac{\mathbf{P}(\text{1 arrival in the interval } [1/4 - x, 1/4] \text{ and no arrivals in the interval } [0, 1/4 - x])}{\mathbf{P}(\text{1 arrival in the interval } [0, 1/4])} \\ &= \frac{\mathbf{P}(\text{1 arrival in the interval } [1/4 - x, 1/4]) \mathbf{P}(\text{no arrivals in the interval } [0, 1/4 - x])}{\mathbf{P}(\text{1 arrival in the interval } [0, 1/4])} \\ &= \frac{\mathbf{P}(1, x) \mathbf{P}(0, 1/4 - x)}{\mathbf{P}(1, 1/4)} \\ &= \frac{e^{-21x} (21x) e^{-21(1/4-x)}}{e^{-21/4} (21/4)} \\ &= \begin{cases} 0, & x < 0, \\ 4x, & x \in [0, 1/4], \\ 1, & x > 1/4. \end{cases} \end{aligned}$$

Thus, we find that the X is uniform over the interval $[0, 1/4]$, with PDF

$$f_X(x) = \begin{cases} 4, & x \in [0, 1/4], \\ 0, & \text{otherwise.} \end{cases}$$

The total time that Al spends on his second lap is $T = X + Y$. Since X and Y correspond to disjoint time intervals in the Poisson process, they are independent, and therefore we can use convolution to find the PDF of T :

$$\begin{aligned} f_T(t) &= \int_{-\infty}^{\infty} f_X(x)f_Y(t-x) dx \\ &= \int_0^{\min(1/4,t)} 4 \cdot 21e^{-21(t-x)} dx \\ &= \begin{cases} 4e^{-21t} (e^{21\min(1/4,t)} - 1), & t \geq 0, \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

Problem 5. (25 points)

(a) (5 points)

Using the law of iterated expectations and the law of total variance,

$$\begin{aligned} \mathbf{E}[N] &= \mathbf{E}[\mathbf{E}[N | X]] \\ &= \mathbf{E}[X] \\ &= \frac{2}{\lambda}, \end{aligned}$$

$$\begin{aligned} \text{var}(N) &= \mathbf{E}[\text{var}(N | X)] + \text{var}(\mathbf{E}[N | X]) \\ &= \mathbf{E}[X] + \text{var}(X) \\ &= \frac{2}{\lambda} + \frac{2}{\lambda^2}, \end{aligned}$$

where $\text{var}(N | X) = \mathbf{E}[N | X] = X$.

(b) (5 points)

$$\begin{aligned} p_N(n) &= \int_x f_X(x)p_{N|X}(n | x)dx \\ &= \int_{x=0}^{\infty} \frac{\lambda^2}{n!} x^{n+1} e^{-(1+\lambda)x} dx \\ &= \frac{\lambda^2}{n!} \cdot \frac{(n+1)!}{(1+\lambda)^{n+2}} \\ &= \begin{cases} \frac{\lambda^2(n+1)}{(1+\lambda)^{n+2}} & n = 0, 1, 2 \dots \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

(c) (5 points)

The equation for $\hat{X}_{\text{lin}}(N)$, the linear least-squares estimator of X based on an observation of N , is

$$\hat{X}_{\text{lin}}(N) = \mathbf{E}[X] + \frac{\text{cov}(X, N)}{\text{var}(N)}(N - \mathbf{E}(N)).$$

The only unknown quantity is $\text{cov}(X, N) = \mathbf{E}[XN] - \mathbf{E}[X]\mathbf{E}[N] = \mathbf{E}[XN] - (\mathbf{E}[X])^2$. Using the law of iterated expectations again,

$$\begin{aligned}\mathbf{E}[XN] &= \mathbf{E}[\mathbf{E}[XN | X]] \\ &= \mathbf{E}[X\mathbf{E}[N | X]] \\ &= \mathbf{E}[X^2] = \text{var}(X) + (\mathbf{E}[X])^2 \\ &= \frac{6}{\lambda^2}.\end{aligned}$$

Thus, $\text{cov}(X, N) = 6/\lambda^2 - 4/\lambda^2 = 2/\lambda^2$. Combining this result with those from (a),

$$\begin{aligned}\hat{X}_{\text{lin}}(N) &= \frac{2}{\lambda} + \frac{\frac{2}{\lambda^2}}{\frac{2}{\lambda} + \frac{2}{\lambda^2}} \left(N - \frac{2}{\lambda} \right) \\ &= \frac{2+N}{1+\lambda}.\end{aligned}$$

(d) (5 points)

The expression for $\hat{X}_{\text{MAP}}(N)$, the MAP estimator of X based on an observation of N is

$$\begin{aligned}\hat{X}_{\text{MAP}}(N) &= \arg \max_x f_{X|N}(x | n) \\ &= \arg \max_x \frac{f_X(x)p_{N|X}(n | x)}{p_N(n)} \\ &= \arg \max_x f_X(x)p_{N|X}(n | x) \\ &= \arg \max_x \frac{\lambda^2}{n!} x^{n+1} e^{-(1+\lambda)x} \\ &= \arg \max_x x^{n+1} e^{-(1+\lambda)x},\end{aligned}$$

where the third equality holds since $p_N(n)$ has no dependency on x and the last equality holds by removing all quantities that have no dependency on x . The max can be found by differentiation and the result is:

$$\hat{X}_{\text{MAP}}(N) = \frac{1+N}{1+\lambda}.$$

This is the only local extremum in the range $x \in [0, \infty)$. Moreover, $f_{X|N}(x | n)$ equals 0 at $x = 0$ and goes to 0 as $x \rightarrow \infty$ and $f_{X|N}(x | n) > 0$ otherwise. We can therefore conclude that $\hat{X}_{\text{MAP}}(N)$ is indeed a maximum.

(e) (5 points)

To minimize the probability of error, we choose the hypothesis that has the larger posterior

probability. We will choose the hypothesis that $X = 2$ if

$$\begin{aligned} \mathbf{P}(X = 2 | N = 3) &> \mathbf{P}(X = 3 | N = 3) \\ \frac{\mathbf{P}(X = 2)\mathbf{P}(N = 3 | X = 2)}{\mathbf{P}(N = 3)} &> \frac{\mathbf{P}(X = 3)\mathbf{P}(N = 3 | X = 3)}{\mathbf{P}(N = 3)} \\ \mathbf{P}(X = 2)\mathbf{P}(N = 3 | X = 2) &> \mathbf{P}(X = 3)\mathbf{P}(N = 3 | X = 3) \\ \frac{3^3}{35} \cdot \frac{2^3 e^{-2}}{3!} &> \frac{2^3}{35} \cdot \frac{3^3 e^{-3}}{3!} \\ e^{-2} &> e^{-3}. \end{aligned}$$

The inequality holds so we choose the hypothesis that $X = 2$ to minimize the probability of error.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Final Exam | Spring 2009)

Problem 1: True or False (2pts. each, 18 pts. total)

No partial credit will be given for individual questions in this part of the quiz.

- a. Let $\{X_n\}$ be a sequence of i.i.d random variables taking values in the interval $[0, 0.5]$. Consider the following statements:
- (A) If $\mathbf{E}[X_n^2]$ converges to 0 as $n \rightarrow \infty$ then X_n converges to 0 in probability.
 - (B) If all X_n have $\mathbf{E}[X_n] = 0.2$ and $\text{var}(X_n)$ converges to 0 as $n \rightarrow \infty$ then X_n converges to 0.2 in probability.
 - (C) The sequence of random variables Z_n , defined by $Z_n = X_1 \cdot X_2 \cdots X_n$, converges to 0 in probability as $n \rightarrow \infty$.

Which of these statements are **always** true? Write **True** or **False** in each of the boxes below.

A:	B:	C:
----	----	----

- b. Let X_i ($i = 1, 2, \dots$) be i.i.d. random variables with mean 0 and variance 2; Y_i ($i = 1, 2, \dots$) be i.i.d. random variables with mean 2. Assume that all variables X_i, Y_j are independent. Consider the following statements:

- (A) $\frac{X_1 + \dots + X_n}{n}$ converges to 0 in probability as $n \rightarrow \infty$.
- (B) $\frac{X_1^2 + \dots + X_n^2}{n}$ converges to 2 in probability as $n \rightarrow \infty$.
- (C) $\frac{X_1 Y_1 + \dots + X_n Y_n}{n}$ converges to 0 in probability as $n \rightarrow \infty$.

Which of these statements are **always** true? Write **True** or **False** in each of the boxes below.

A:	B:	C:
----	----	----

- c. We have i.i.d. random variables $X_1 \dots X_n$ with an unknown distribution, and with $\mu = \mathbf{E}[X_i]$. We define $M_n = (X_1 + \dots + X_n)/n$. Consider the following statements:

- (A) M_n is a maximum-likelihood estimator for μ , irrespective of the distribution of the X_i 's.
- (B) M_n is a consistent estimator for μ , irrespective of the distribution of the X_i 's.
- (C) M_n is an asymptotically unbiased estimator for μ , irrespective of the distribution of the X_i 's.

Which of these statements are **always** true? Write **True** or **False** in each of the boxes below.

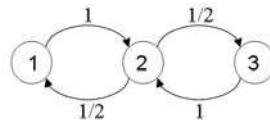
A:	B:	C:
----	----	----

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Final Exam | Spring 2009)

Problem 2: Multiple Choice (4 pts. each, 24 pts. total)

Clearly circle the appropriate choice. **No partial credit** will be given for individual questions in this part of the quiz.

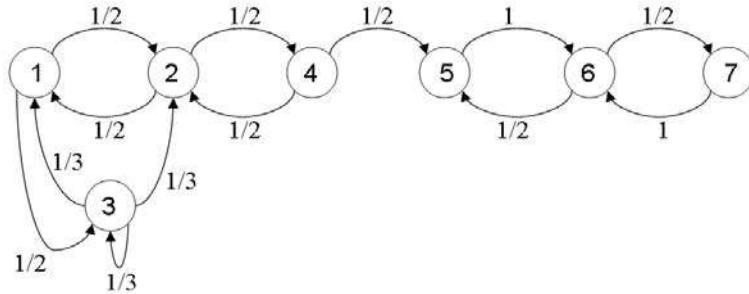
- a. Earthquakes in Sumatra occur according to a Poisson process of rate $\lambda = 2/\text{year}$. Conditioned on the event that exactly two earthquakes take place in a year, what is the probability that both earthquakes occur in the first three months of the year? (for simplicity, assume all months have 30 days, and each year has 12 months, i.e., 360 days).
- (i) $1/12$
 - (ii) $1/16$
 - (iii) $64/225$
 - (iv) $4e^{-4}$
 - (v) There is not enough information to determine the required probability.
 - (vi) None of the above.
- b. Consider a continuous-time Markov chain with three states $i \in \{1, 2, 3\}$, with dwelling time in each visit to state i being an exponential random variable with parameter $\nu_i = i$, and transition probabilities p_{ij} defined by the graph



What is the long-term expected fraction of time spent in state 2?

- (i) $1/2$
- (ii) $1/4$
- (iii) $2/5$
- (iv) $3/7$
- (v) None of the above.

c. Consider the following Markov chain:



Starting in state 3, what is the steady-state probability of being in state 1?

- (i) 1/3
- (ii) 1/4
- (iii) 1
- (iv) 0
- (v) None of the above.

d. Random variables X and Y are such that the pair (X, Y) is uniformly distributed over the trapezoid A with corners $(0,0)$, $(1,2)$, $(3,2)$, and $(4,0)$ shown in Fig. 1:

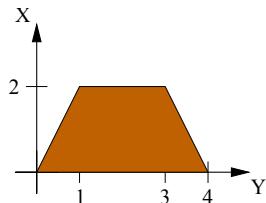


Figure 1: $f_{X,Y}(x,y)$ is constant over the shaded area, zero otherwise.

i.e.

$$f_{X,Y}(x,y) = \begin{cases} c, & (x,y) \in A \\ 0, & \text{else.} \end{cases}$$

We observe Y and use it to estimate X . Let \hat{X} be the least mean squared error estimator of X given Y . What is the value of $\text{var}(\hat{X} - X|Y = 1)$?

- (i) 1/6
- (ii) 3/2
- (iii) 1/3
- (iv) The information is not sufficient to compute this value.
- (v) None of the above.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Final Exam | Spring 2009)

e. $X_1 \dots X_n$ are i.i.d. normal random variables with mean value μ and variance v . Both μ and v are unknown. We define $M_n = (X_1 + \dots + X_n)/n$ and

$$V_n = \frac{1}{n-1} \sum_{i=1}^n (X_i - M_n)^2$$

We also define $\Phi(x)$ to be the CDF for the standard normal distribution, and $\Psi_{n-1}(x)$ to be the CDF for the t-distribution with $n-1$ degrees of freedom. Which of the following choices gives an exact 99% confidence interval for μ for all $n > 1$?

- (i) $[M_n - \delta \sqrt{\frac{V_n}{n}}, M_n + \delta \sqrt{\frac{V_n}{n}}]$ where δ is chosen to give $\Phi(\delta) = 0.99$.
- (ii) $[M_n - \delta \sqrt{\frac{V_n}{n}}, M_n + \delta \sqrt{\frac{V_n}{n}}]$ where δ is chosen to give $\Phi(\delta) = 0.995$.
- (iii) $[M_n - \delta \sqrt{\frac{V_n}{n}}, M_n + \delta \sqrt{\frac{V_n}{n}}]$ where δ is chosen to give $\Psi_{n-1}(\delta) = 0.99$.
- (iv) $[M_n - \delta \sqrt{\frac{V_n}{n}}, M_n + \delta \sqrt{\frac{V_n}{n}}]$ where δ is chosen to give $\Psi_{n-1}(\delta) = 0.995$.
- (v) None of the above.

f. We have i.i.d. random variables X_1, X_2 which have an exponential distribution with unknown parameter θ . Under hypothesis $H_0, \theta = 1$. Under hypothesis $H_1, \theta = 2$. Under a likelihood-ratio test, the rejection region takes which of the following forms?

- (i) $R = \{(x_1, x_2) : x_1 + x_2 > \xi\}$ for some value ξ .
- (ii) $R = \{(x_1, x_2) : x_1 + x_2 < \xi\}$ for some value ξ .
- (iii) $R = \{(x_1, x_2) : e^{x_1} + e^{x_2} > \xi\}$ for some value ξ .
- (iv) $R = \{(x_1, x_2) : e^{x_1} + e^{x_2} < \xi\}$ for some value ξ .
- (v) None of the above.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Final Exam | Spring 2009)

Problem 3 (12 pts. total)

Aliens of two races (blue and green) are arriving on Earth independently according to Poisson process distributions with parameters λ_b and λ_a respectively. The Alien Arrival Registration Service Authority (AARSA) will begin registering alien arrivals soon.

Let T_1 denote the time AARSA will function until it registers its first alien. Let G be the event that the first alien to be registered is a green one. Let T_2 be the time AARSA will function until at least one alien of both races is registered.

- (a) (4 points.) Express $\mu_1 = \mathbf{E}[T_1]$ in terms of λ_g and λ_b . **Show your work.**

- (b) (4 points.) Express $p = \mathbf{P}(G)$ in terms of λ_g and λ_b . **Show your work.**

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Final Exam | Spring 2009)

(c) (4 points.) Express $\mu_2 = \mathbf{E}[T_2]$ in terms of λ_g and λ_b . **Show your work.**

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Final Exam | Spring 2009)

Problem 4 (18 pts. total)

Researcher Jill is interested in studying employment in technology firms in Dilicon Valley. She denotes by X_i the number of employees in technology firm i and assumes that X_i are independent and identically distributed with mean p . To estimate p , Jill randomly interviews n technology firms and observes the number of employees in these firms.

- (a) (6 points.) Jill uses

$$M_n = \frac{X_1 + \cdots + X_n}{n}$$

as an estimator for p . Find the limit of $\mathbf{P}(M_n \leq x)$ as $n \rightarrow \infty$ for $x < p$. Find the limit of $\mathbf{P}(M_n \leq x)$ as $n \rightarrow \infty$ for $x > p$. **Show your work.**

- (b) (6 points.) Find the smallest n , the number of technology firms Jill must sample, for which the Chebyshev inequality yields a guarantee

$$\mathbf{P}(|M_n - p| \geq 0.5) \leq 0.05.$$

Assume that $\text{var}(X_i) = v$ for some constant v . State your solution as a function of v . **Show your work.**

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Final Exam | Spring 2009)

- (c) (6 points.) Assume now that the researcher samples $n = 5000$ firms. Find an approximate value for the probability

$$\mathbf{P}(|M_{5000} - p| \geq 0.5)$$

using the Central Limit Theorem. Assume again that $\text{var}(X_i) = v$ for some constant v . Give your answer in terms of v , and the standard normal CDF Φ . **Show your work.**

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Final Exam | Spring 2009)

Problem 5 (12 pts. total)

The RandomView window factory produces window panes. After manufacturing, 1000 panes were loaded onto a truck. The weight W_i of the i -th pane (in pounds) on the truck is modeled as a random variable, with the assumption that the W_i 's are independent and identically distributed.

- (a) (6 points.) Assume that the measured weight of the load on the truck was 2340 pounds, and that $\text{var}(W_i) \leq 4$. Find an approximate 95 percent confidence interval for $\mu = \mathbf{E}[W_i]$, using the Central Limit Theorem (you may use the standard normal table which was handed out with this quiz). **Show your work.**

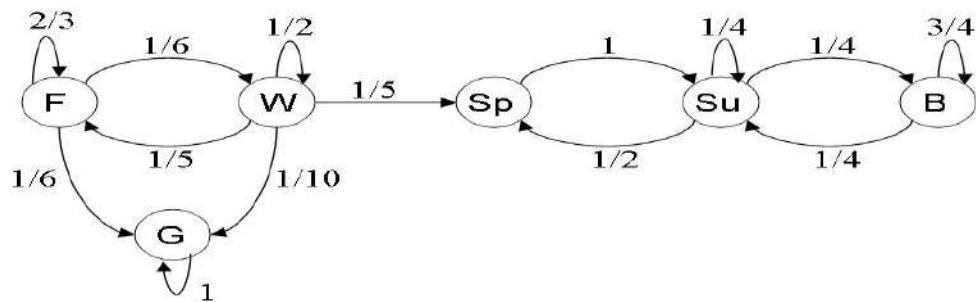
- (b) (6 points.) Now assume instead that the random variables W_i are i.i.d, with an exponential distribution with parameter $\theta > 0$, i.e., a distribution with PDF

$$f_W(w; \theta) = \theta e^{-\theta w}$$

What is the maximum likelihood estimate of θ , given that the truckload has weight 2340 pounds?
Show your work.

Problem 6 (21 pts. total)

In Alice's Wonderland, there are six different seasons: Fall (F), Winter (W), Spring (Sp), Summer (Su), Bitter Cold (B), and Golden Sunshine (G). The seasons do not follow any particular order, instead, at the beginning of each day the Head Wizard assigns the season for the day, according to the following Markov chain model:



Thus, for example, if it is Fall one day then there is $1/6$ probability that it will be Winter the next day (note that it is possible to have the same season again the next day).

- (a) (4 points.) For each state in the above chain, identify whether it is recurrent or transient. **Show your work.**

(b) (4 points.) If it is Fall on Monday, what is the probability that it will be Summer on Thursday of the same week? **Show your work.**

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Final Exam | Spring 2009)

- (c) (4 points.) If it is Spring today, will the chain converge to steady-state probabilities? If so, compute the steady-state probability for each state. If not, explain why these probabilities do not exist. **Show your work.**
- (d) (5 points.) If it is Fall today, what is the probability that Bitter Cold will never arrive in the future? **Show your work.**

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Final Exam | Spring 2009)

- (e) (5 points.) If it is Fall today, what is the expected number of days till either Summer or Golden Sunshine arrives for the first time? **Show your work.**

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Final Exam | Spring 2009)

Problem 7 (12 pts. total)

A newscast covering the final baseball game between Sed Rox and Y Nakee becomes noisy at the crucial moment when the viewers are informed whether Y Nakee won the game.

Let a be the parameter describing the actual outcome: $a = 1$ if Y Nakee won, $a = -1$ otherwise. There were n viewers listening to the telecast. Let Y_i be the information received by viewer i ($1 \leq i \leq n$). Under the noisy telecast, $Y_i = a$ with probability p , and $Y_i = -a$ with probability $1 - p$. Assume that the random variables Y_i are independent of each other.

The viewers as a group come up with a joint estimator

$$Z_n = \begin{cases} 1 & \text{if } \sum_{i=1}^n Y_i \geq 0, \\ -1 & \text{otherwise.} \end{cases}$$

(a) (6 points.)

Find $\lim_{n \rightarrow \infty} \mathbf{P}(Z_n = a)$ assuming that $p > 0.5$ and $a = 1$. **Show your work.**

(b) (6 points.) Find $\lim_{n \rightarrow \infty} \mathbf{P}(Z_n = a)$, assuming that $p = 0.5$ and $a = 1$. **Show your work.**

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Final Exam Solutions | Spring 2009)

Problem 1: True or False (2pts. each, 18 pts. total)

No partial credit will be given for individual questions in this part of the quiz.

- a. Let $\{X_n\}$ be a sequence of i.i.d. random variables taking values in the interval $[0, 0.5]$. Consider the following statements:
- (A) If $\mathbf{E}[X_n^2]$ converges to 0 as $n \rightarrow \infty$ then X_n converges to 0 in probability.
 - (B) If all X_n have $\mathbf{E}[X_n] = 0.2$ and $\text{var}(X_n)$ converges to 0 as $n \rightarrow \infty$ then X_n converges to 0.2 in probability.
 - (C) The sequence of random variables Z_n , defined by $Z_n = X_1 \cdot X_2 \cdots X_n$, converges to 0 in probability as $n \rightarrow \infty$.

Which of these statements are **always** true? Write **True** or **False** in each of the boxes below.

 A: True

 B: True

 C: True

Solution:

- (A) True. The fact that $\lim_{n \rightarrow \infty} \mathbf{E}[X_n^2] = 0$ implies $\lim_{n \rightarrow \infty} \mathbf{E}[X_n] = 0$ and $\lim_{n \rightarrow \infty} \text{var}(X_n) = 0$. Hence, one has

$$\begin{aligned}\mathbf{P}(|X_n - 0| \geq \epsilon) &\leq \mathbf{P}(|X_n - \mathbf{E}[X_n]| \geq \epsilon/2) + \mathbf{P}(|\mathbf{E}[X_n] - 0| \geq \epsilon/2) \\ &\leq \frac{\text{var}(X_n)}{(\epsilon/2)^2} + \mathbf{P}(|\mathbf{E}[X_n] - 0| \geq \epsilon/2) \rightarrow 0,\end{aligned}$$

where we have applied Chebyshev inequality.

- (B) True. Applying Chebyshev inequality gives

$$\mathbf{P}(|X_n - \mathbf{E}[X_n]| \geq \epsilon) \leq \frac{\text{var}(X_n)}{\epsilon^2} \rightarrow 0.$$

Hence X_n converges to $\mathbf{E}[X_n] = 0.2$ in probability.

- (C) True. For all $\epsilon > 0$, since $Z_n \leq (1/2)^n \Rightarrow \mathbf{P}(|Z_n - 0| \geq \epsilon) = 0$ for $n > -\log \epsilon / \log 2$.

- b. Let X_i ($i = 1, 2, \dots$) be i.i.d. random variables with mean 0 and variance 2; Y_i ($i = 1, 2, \dots$) be i.i.d. random variables with mean 2. Assume that all variables X_i, Y_j are independent. Consider the following statements:

- (A) $\frac{X_1 + \dots + X_n}{n}$ converges to 0 in probability as $n \rightarrow \infty$.
- (B) $\frac{X_1^2 + \dots + X_n^2}{n}$ converges to 2 in probability as $n \rightarrow \infty$.
- (C) $\frac{X_1 Y_1 + \dots + X_n Y_n}{n}$ converges to 0 in probability as $n \rightarrow \infty$.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
 (Final Exam Solutions | Spring 2009)

Which of these statements are **always** true? Write **True** or **False** in each of the boxes below.

A: True	B: True	C: True
---------	---------	---------

Solution:

- (A) True. Note that $\mathbf{E}[\frac{X_1+\dots+X_n}{n}] = 0$ and $\text{var}(\frac{X_1+\dots+X_n}{n}) = \frac{n \cdot 2}{n^2} = \frac{2}{n}$. One can see $\frac{X_1+\dots+X_n}{n}$ converges to 0 in probability.
 - (B) True. Let $Z_i = X_i^2$ and $\mathbf{E}[Z_i] = 2$. Note Z_i are i.i.d. since X_i are i.i.d., and hence one has that $\frac{Z_1+\dots+Z_n}{n}$ converges to $\mathbf{E}[Z_i] = 2$ in probability by the WLLN.
 - (C) True. Let $W_i = X_i Y_i$ and $\mathbf{E}[W_i] = \mathbf{E}[X_i]\mathbf{E}[Y_i] = 0$. Note W_i are i.i.d. since X_i and Y_i are respectively i.i.d., and hence one has that $\frac{W_1+\dots+W_n}{n}$ converges to $\mathbf{E}[W_i] = 0$ in probability by the WLLN.
- c. We have i.i.d. random variables $X_1 \dots X_n$ with an unknown distribution, and with $\mu = \mathbf{E}[X_i]$. We define $M_n = (X_1 + \dots + X_n)/n$. Consider the following statements:
- (A) M_n is a maximum-likelihood estimator for μ , irrespective of the distribution of the X_i 's.
 - (B) M_n is a consistent estimator for μ , irrespective of the distribution of the X_i 's.
 - (C) M_n is an asymptotically unbiased estimator for μ , irrespective of the distribution of the X_i 's.

Which of these statements are **always** true? Write **True** or **False** in each of the boxes below.

A: False	B: True	C: True
----------	---------	---------

Solution:

- (A) False. Consider X_i follow a uniform distribution $U[\mu - \frac{1}{2}, \mu + \frac{1}{2}]$. The ML estimator for μ is any value between $\max(X_1, \dots, X_n) - \frac{1}{2}$ and $\min(X_1, \dots, X_n) + \frac{1}{2}$, instead of M_n .
- (B) True. By the WLLN, M_n converges to μ in probability and hence it is a consistent estimator.
- (C) True. Since $\mathbf{E}[M_n] = \mathbf{E}[X_i] = \mu$, M_n is unbiased estimator for μ and hence asymptotically unbiased.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Final Exam Solutions | Spring 2009)

Problem 2: Multiple Choice (4 pts. each, 24 pts. total)

Clearly circle the appropriate choice. **No partial credit** will be given for individual questions in this part of the quiz.

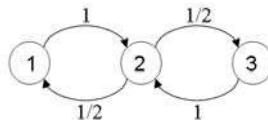
- a. Earthquakes in Sumatra occur according to a Poisson process of rate $\lambda = 2/\text{year}$. Conditioned on the event that exactly two earthquakes take place in a year, what is the probability that both earthquakes occur in the first three months of the year? (for simplicity, assume all months have 30 days, and each year has 12 months, i.e., 360 days).
- (i) $1/12$
 - (ii) 1/16
 - (iii) $64/225$
 - (iv) $4e^{-4}$
 - (v) There is not enough information to determine the required probability.
 - (vi) None of the above.

Solution: Consider the interval of a year be $[0, 1]$.

$$\begin{aligned} \mathbf{P}\left(2 \text{ in } [0, \frac{1}{4}) \mid 2 \text{ in } [0, 1]\right) &= \frac{\mathbf{P}(2 \text{ in } [0, \frac{1}{4}), 0 \text{ in } [\frac{1}{4}, 1])}{\mathbf{P}(2 \text{ in } [0, 1]))} \\ &= \frac{\frac{(\lambda \cdot 1/4)^2}{2!} e^{-\lambda \cdot 1/4} \cdot \frac{(\lambda \cdot 3/4)^0}{0!} e^{-\lambda \cdot 3/4}}{\frac{\lambda^2}{2!} e^{-\lambda}} \\ &= \frac{1}{16} \end{aligned}$$

(alternative explanation) Given that exactly two earthquakes happened in 12 months, each earthquake is equally likely to happen in any month of the 12, the probability that it happens in the first 3 months is $3/12 = 1/4$. The probability that both happen in the first 3 months is $(1/4)^2$.

- b. Consider a continuous-time Markov chain with three states $i \in \{1, 2, 3\}$, with dwelling time in each visit to state i being an exponential random variable with parameter $\nu_i = i$, and transition probabilities p_{ij} defined by the graph



What is the long-term expected fraction of time spent in state 2?

- (i) $1/2$
- (ii) $1/4$
- (iii) $2/5$

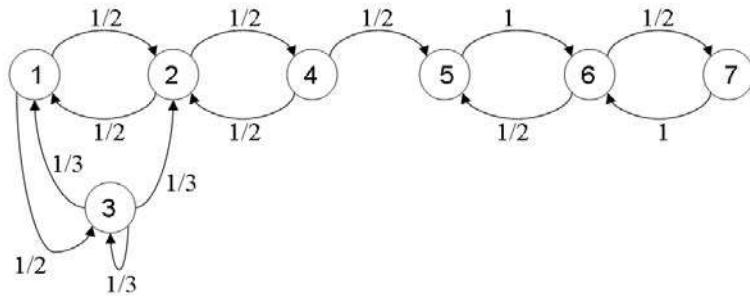
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Final Exam Solutions | Spring 2009)

(iv) 3/7

(v) None of the above.

Solution: First, we calculate the $q_{ij} = \nu_i p_{ij}$, i.e., $q_{12} = q_{21} = q_{23} = 1$ and $q_{32} = 3$. The balance and normalization equations of this birth-death markov chain can be expressed as, $\pi_1 = \pi_2$, $\pi_2 = 3\pi_3$ and $\pi_1 + \pi_2 + \pi_3 = 1$, yielding $\pi_2 = 3/7$.

c. Consider the following Markov chain:



Starting in state 3, what is the steady-state probability of being in state 1?

- (i) $1/3$
- (ii) $1/4$
- (iii) 1
- (iv) $\boxed{0}$
- (v) None of the above.

Solution: State 1 is transient.

d. Random variables X and Y are such that the pair (X, Y) is uniformly distributed over the trapezoid A with corners $(0, 0)$, $(1, 2)$, $(3, 2)$, and $(4, 0)$ shown in Fig. 1:

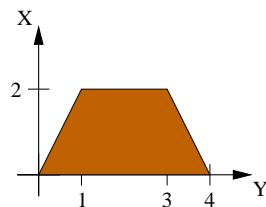


Figure 1: $f_{X,Y}(x,y)$ is constant over the shaded area, zero otherwise.

i.e.

$$f_{X,Y}(x,y) = \begin{cases} c, & (x,y) \in A \\ 0, & \text{else.} \end{cases}$$

We observe Y and use it to estimate X . Let \hat{X} be the least mean squared error estimator of X given Y . What is the value of $\text{var}(\hat{X} - X|Y = 1)$?

- (i) $1/6$
- (ii) $3/2$
- (iii) $\boxed{1/3}$
- (iv) The information is not sufficient to compute this value.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Final Exam Solutions | Spring 2009)

- (v) None of the above.

Solution: $f_{X|Y=1}(x)$ is uniform on $[0, 2]$ therefore $\hat{X} = \mathbf{E}[X|Y = 1] = 1$ and $\text{var}(\hat{X} - X|Y = 1) = \text{var}(X|Y = 1) = (2 - 0)^2/12 = 1/3$.

- e. $X_1 \dots X_n$ are i.i.d. normal random variables with mean value μ and variance v . Both μ and v are unknown. We define $M_n = (X_1 + \dots + X_n)/n$ and

$$V_n = \frac{1}{n-1} \sum_{i=1}^n (X_i - M_n)^2$$

We also define $\Phi(x)$ to be the CDF for the standard normal distribution, and $\Psi_{n-1}(x)$ to be the CDF for the t-distribution with $n - 1$ degrees of freedom. Which of the following choices gives an exact 99% confidence interval for μ for all $n > 1$?

- (i) $[M_n - \delta \sqrt{\frac{V_n}{n}}, M_n + \delta \sqrt{\frac{V_n}{n}}]$ where δ is chosen to give $\Phi(\delta) = 0.99$.
- (ii) $[M_n - \delta \sqrt{\frac{V_n}{n}}, M_n + \delta \sqrt{\frac{V_n}{n}}]$ where δ is chosen to give $\Phi(\delta) = 0.995$.
- (iii) $[M_n - \delta \sqrt{\frac{V_n}{n}}, M_n + \delta \sqrt{\frac{V_n}{n}}]$ where δ is chosen to give $\Psi_{n-1}(\delta) = 0.99$.
- (iv) $[M_n - \delta \sqrt{\frac{V_n}{n}}, M_n + \delta \sqrt{\frac{V_n}{n}}]$ where δ is chosen to give $\Psi_{n-1}(\delta) = 0.995$.
- (v) None of the above.

Solution: See Lecture 23, slides 10-12.

- f. We have i.i.d. random variables X_1, X_2 which have an exponential distribution with unknown parameter θ . Under hypothesis H_0 , $\theta = 1$. Under hypothesis H_1 , $\theta = 2$. Under a likelihood-ratio test, the rejection region takes which of the following forms?

- (i) $R = \{(x_1, x_2) : x_1 + x_2 > \xi\}$ for some value ξ .
- (ii) $R = \{(x_1, x_2) : x_1 + x_2 < \xi\}$ for some value ξ .
- (iii) $R = \{(x_1, x_2) : e^{x_1} + e^{x_2} > \xi\}$ for some value ξ .
- (iv) $R = \{(x_1, x_2) : e^{x_1} + e^{x_2} < \xi\}$ for some value ξ .
- (v) None of the above.

Solution: We defined $R = \{x = (x_1, x_2) | L(x) > c\}$ where

$$L(x) = \frac{f_X(x; H_1)}{f_X(x; H_0)} = \frac{\theta_1 e^{-\theta_1 x_1} \theta_1 e^{-\theta_1 x_2}}{\theta_0 e^{-\theta_0 x_1} \theta_0 e^{-\theta_0 x_2}} = \frac{\theta_1^2}{\theta_0^2} e^{(\theta_0 - \theta_1)(x_1 + x_2)} = 4e^{-(x_1 + x_2)}.$$

So $R = \{(x_1, x_2) | x_1 + x_2 < -\log(c/4)\}$

Problem 3 (12 pts. total)

Aliens of two races (blue and green) are arriving on Earth independently according to Poisson process distributions with parameters λ_b and λ_g respectively. The Alien Arrival Registration Service Authority (AARSA) will begin registering alien arrivals soon.

Let T_1 denote the time AARSA will function until it registers its first alien. Let G be the event that the first alien to be registered is a green one. Let T_2 be the time AARSA will function until at least one alien of both races is registered.

- (a) (4 points.) Express $\mu_1 = \mathbf{E}[T_1]$ in terms of λ_g and λ_b . **Show your work.**

Answer: $\mu_1 = \mathbf{E}[T_1] = \frac{1}{\lambda_g + \lambda_b}$

Solution: We consider the process of arrivals of both types of Aliens. This is a merged Poisson process with arrival rate $\lambda_g + \lambda_b$. T_1 is the time until the first arrival, and therefore is exponentially distributed with parameter $\lambda_g + \lambda_b$. Therefore $\mu_1 = \mathbf{E}[T_1] = \frac{1}{\lambda_g + \lambda_b}$.

One can also go about this using derived distributions, since $T_1 = \min(T_1^g, T_1^b)$ where T_1^g and T_1^b are the first arrival times of green and blue Aliens respectively (i.e., T_1^g and T_1^b are exponentially distributed with parameters λ_g and λ_b , respectively.)

- (b) (4 points.) Express $p = \mathbf{P}(G)$ in terms of λ_g and λ_b . **Show your work.**

Answer: $\mathbf{P}(G) = \frac{\lambda_g}{\lambda_g + \lambda_b}$

Solution: We consider the same merged Poisson process as before, with arrival rate $\lambda_g + \lambda_b$. Any particular arrival of the merged process has probability $\frac{\lambda_g}{\lambda_g + \lambda_b}$ of corresponding to a green Alien and probability $\frac{\lambda_b}{\lambda_g + \lambda_b}$ of corresponding to a blue Alien. The question asks for $\mathbf{P}(G) = \frac{\lambda_g}{\lambda_g + \lambda_b}$.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Final Exam Solutions | Spring 2009)

(c) (4 points.) Express $\mu_2 = \mathbf{E}[T_2]$ in terms of λ_g and λ_b .

Show your work.

Answer: $\frac{1}{\lambda_g + \lambda_b} + \frac{\lambda_g}{\lambda_g + \lambda_b} \left(\frac{1}{\lambda_b} \right) + \frac{\lambda_b}{\lambda_g + \lambda_b} \left(\frac{1}{\lambda_g} \right)$

Solution: The time T_2 until at least one green and one red Aliens have arrived can be expressed as $T_2 = \max(T_1^g, T_1^b)$, where T_1^g and T_1^b are the first arrival times of green and blue Aliens respectively (i.e., T_1^g and T_1^b are exponentially distributed with parameters λ_g and λ_b , respectively.)

The expected time till the 1st Alien arrives was calculated in (a), $\mu_1 = \mathbf{E}[T_1] = \frac{1}{\lambda_g + \lambda_b}$. To compute the remaining time we simply condition on the 1st Alien being green(e.g. event G) or blue(event G^c), and use the memoryless property of Poisson, i.e.,

$$\begin{aligned}
 \mathbf{E}[T_2] &= \mathbf{E}[T_1] + \mathbf{P}(G)\mathbf{E}[\text{Time until first Blue arrives}|G] + \mathbf{P}(G^c)\mathbf{E}[\text{Time until first Green arrives}|G^c] \\
 &= \mathbf{E}[T_1] + \mathbf{P}(G)\mathbf{E}[T_2^b] + (1 - \mathbf{P}(G))\mathbf{E}[T_2^g] \\
 &= \frac{1}{\lambda_g + \lambda_b} + \frac{\lambda_g}{\lambda_g + \lambda_b} \left(\frac{1}{\lambda_b} \right) + \frac{\lambda_b}{\lambda_g + \lambda_b} \left(\frac{1}{\lambda_g} \right)
 \end{aligned}$$

Problem 4 (18 pts. total)

Researcher Jill is interested in studying employment in technology firms in Silicon Valley. She denotes by X_i the number of employees in technology firm i and assumes that X_i are independent and identically distributed with mean p . To estimate p , Jill randomly interviews n technology firms and observes the number of employees in these firms.

- (a) (6 points.) Jill uses

$$M_n = \frac{X_1 + \cdots + X_n}{n}$$

as an estimator for p . Find the limit of $\mathbf{P}(M_n \leq x)$ as $n \rightarrow \infty$ for $x < p$. Find the limit of $\mathbf{P}(M_n \leq x)$ as $n \rightarrow \infty$ for $x > p$. **Show your work.**

Solution: Since X_i is i.i.d., M_n converges to p in probability, i.e., $\lim_{n \rightarrow \infty} \mathbf{P}(|M_n - p| > \epsilon) = 0$, implying $\lim_{n \rightarrow \infty} \mathbf{P}(M_n < p - \epsilon) = 0$ and $\lim_{n \rightarrow \infty} \mathbf{P}(M_n > p + \epsilon) = 0$, for all $\epsilon > 0$. Hence

$$\lim_{n \rightarrow \infty} \mathbf{P}(M_n \leq x) = \begin{cases} 0, & x < p, \\ 1, & x > p. \end{cases}$$

- (b) (6 points.) Find the smallest n , the number of technology firms Jill must sample, for which the Chebyshev inequality yields a guarantee

$$\mathbf{P}(|M_n - p| \geq 0.5) \leq 0.05.$$

Assume that $\text{var}(X_i) = v$ for some constant v . State your solution as a function of v . **Show your work.**

Solution: Since M_n converges to p in probability and $\text{var}(M_n) = \frac{n}{n^2} \cdot \text{var}(X_i) = v/n$, Chebyshev inequality gives

$$\mathbf{P}(|M_n - p| \geq 0.5) \leq \frac{\text{var}(M_n)}{0.5^2} = \frac{v}{n \cdot 0.5^2} = 0.05$$

$$\Rightarrow n = 80v.$$

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Final Exam Solutions | Spring 2009)

- (c) (6 points.) Assume now that the researcher samples $n = 5000$ firms. Find an approximate value for the probability

$$\mathbf{P}(|M_{5000} - p| \geq 0.5)$$

using the Central Limit Theorem. Assume again that $\text{var}(X_i) = v$ for some constant v . Give your answer in terms of v , and the standard normal CDF Φ . **Show your work.**

Solution: By CLT, we can approximate by a standard normal distribution

$$\frac{\sum_{i=1}^n X_i - np}{\sqrt{nv}}$$

when n is large, and hence,

$$\mathbf{P}(|M_{5000} - p| \geq 0.5) = P\left(\left|\frac{\sum_{i=1}^n X_i - np}{\sqrt{nv}}\right| \geq \frac{0.5\sqrt{n}}{\sqrt{v}}\right) = \boxed{2 - 2\Phi\left(\frac{0.5\sqrt{n}}{\sqrt{v}}\right)},$$

where $n = 5000$.

Problem 5 (12 pts. total)

The RandomView window factory produces window panes. After manufacturing, 1000 panes were loaded onto a truck. The weight W_i of the i -th pane (in pounds) on the truck is modeled as a random variable, with the assumption that the W_i 's are independent and identically distributed.

- (a) (6 points.) Assume that the measured weight of the load on the truck was 2340 pounds, and that $\text{var}(W_i) \leq 4$. Find an approximate 95 percent confidence interval for $\mu = \mathbf{E}[W_i]$, using the Central Limit Theorem (you may use the standard normal table which was handed out with this quiz). **Show your work.**

Answer: [2.216, 2.464]

Solution: The sample mean estimator $\hat{\Theta}_n = \frac{W_1 + \dots + W_n}{n}$ in this case is

$$\hat{\Theta}_{1000} = \frac{2340}{1000} = 2.34$$

Using the CDF $\Phi(z)$ of the standard normal available in the normal tables, we have $\Phi(1.96) = 0.975$, so we obtain

$$\mathbf{P}\left(\frac{|\hat{\Theta}_{1000} - \mu|}{\sqrt{\text{var}(W_i)/1000}} \leq 1.96\right) \approx 0.95.$$

Because the variance is less than 4, we have

$$\mathbf{P}(|\hat{\Theta}_{1000} - \mu| \leq 1.96\sqrt{\text{var}(W_i)/1000}) \leq \mathbf{P}(|\hat{\Theta}_{1000} - \mu| \leq 1.96\sqrt{4/1000}),$$

and letting the right-hand side of the above equation ≈ 0.95 gives a 95% confidence, i.e.,

$$[\hat{\Theta}_{1000} - 1.96\sqrt{\frac{4}{1000}}, \hat{\Theta}_{1000} + 1.96\sqrt{\frac{4}{1000}}] = [\hat{\Theta}_{1000} - 0.124, \hat{\Theta}_{1000} + 0.124]$$

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Final Exam Solutions | Spring 2009)

- (b) (6 points.) Now assume instead that the random variables W_i are i.i.d, with an exponential distribution with parameter $\theta > 0$, i.e., a distribution with PDF

$$f_W(w; \theta) = \theta e^{-\theta w}$$

What is the maximum likelihood estimate of θ , given that the truckload has weight 2340 pounds?
Show your work.

Answer: $\hat{\Theta}_{1000}^{mle} = \frac{1000}{2340} = 0.4274$

Solution: The likelihood function is

$$f_W(w; \theta) = \prod_{i=1}^n f_{W_i}(w_i; \theta) = \prod_{i=1}^n \theta e^{-\theta w_i},$$

And the log-likelihood function is

$$\log f_W(w; \theta) = n \log \theta - \theta \sum_{i=1}^n w_i,$$

The derivative with respect to θ is $\frac{n}{\theta} - \sum_{i=1}^n w_i$, and by setting it to zero, we see that the maximum of $\log f_W(w; \theta)$ over $\theta \geq 0$ is attained at $\hat{\theta}_n = \frac{n}{\sum_{i=1}^n w_i}$. The resulting estimator is

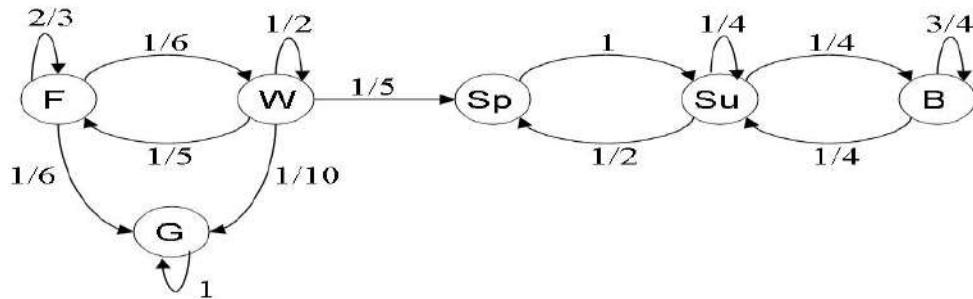
$$\hat{\Theta}_n^{mle} = \frac{n}{\sum_{i=1}^n W_i}.$$

In our case,

$$\hat{\Theta}_{1000}^{mle} = \frac{1000}{2340} = 0.4274$$

Problem 6 (21 pts. total)

In Alice's Wonderland, there are six different seasons: Fall (F), Winter (W), Spring (Sp), Summer (Su), Bitter Cold (B), and Golden Sunshine (G). The seasons do not follow any particular order, instead, at the beginning of each day the Head Wizard assigns the season for the day, according to the following Markov chain model:



Thus, for example, if it is Fall one day then there is $1/6$ probability that it will be Winter the next day (note that it is possible to have the same season again the next day).

- (a) (4 points.) For each state in the above chain, identify whether it is recurrent or transient. **Show your work.**

Solution: F and W are transient states; Sp, Su, B, and G are recurrent states.

- (b) (4 points.) If it is Fall on Monday, what is the probability that it will be Summer on Thursday of the same week? **Show your work.**

Solution: There is only one path from F to Su in three days.

$$\begin{aligned}
 \mathbf{P}(S_4 = \text{Su} | S_1 = \text{F}) &= \mathbf{P}(S_2 = \text{W} | S_1 = \text{F}) \cdot \mathbf{P}(S_3 = \text{Sp} | S_2 = \text{W}) \cdot \mathbf{P}(S_4 = \text{Su} | S_3 = \text{Sp}) \\
 &= \frac{1}{6} \cdot \frac{1}{5} \cdot 1 = \boxed{\frac{1}{30}}
 \end{aligned}$$

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Final Exam Solutions | Spring 2009)

- (c) (4 points.) If it is Spring today, will the chain converge to steady-state probabilities? If so, compute the steady-state probability for each state. If not, explain why these probabilities do not exist. **Show your work.**

Solution: The Markov chain will stay in the recurrent class $\{\text{Sp}, \text{Su}, \text{B}\}$, and

$$\left\{ \begin{array}{l} \pi_{\text{Sp}} \cdot 1 = \pi_{\text{Su}} \cdot \frac{1}{2} \\ \pi_{\text{B}} \cdot \frac{1}{4} = \pi_{\text{Su}} \cdot \frac{1}{4} \\ \pi_{\text{F}} = 0 \\ \pi_{\text{W}} = 0 \\ \pi_{\text{G}} = 0 \\ \pi_{\text{F}} + \pi_{\text{W}} + \pi_{\text{G}} + \pi_{\text{Sp}} + \pi_{\text{Su}} + \pi_{\text{B}} = 1 \end{array} \right.$$

$$\Rightarrow \boxed{\pi_{\text{F}} = 0, \pi_{\text{W}} = 0, \pi_{\text{G}} = 0, \pi_{\text{Sp}} = 1/5, \pi_{\text{Su}} = 2/5, \pi_{\text{B}} = 2/5.}$$

- (d) (5 points.) If it is Fall today, what is the probability that Bitter Cold will never arrive in the future? **Show your work.**

Solution: Let a_F and a_W be the probabilities that Bitter Cold will never arrive starting from Fall and Winter, respectively. This is equivalent to the Markov chain ends up in G.

$$\left\{ \begin{array}{l} a_F = \frac{2}{3} \cdot a_F + \frac{1}{6} \cdot a_W + \frac{1}{6} \cdot 1 \\ a_W = \frac{1}{5} \cdot a_F + \frac{1}{2} \cdot a_W + \frac{1}{10} \cdot 1 \end{array} \right.$$

$$\Rightarrow \boxed{a_F = 3/4.}$$

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Final Exam Solutions | Spring 2009)

- (e) (5 points.) If it is Fall today, what is the expected number of days till either Summer or Golden Sunshine arrives for the first time? **Show your work.**

Solution: Let μ_F and μ_W be expected number of days till either Summer or Golden Sunshine arrives for the first time, respectively.

$$\begin{cases} \mu_F = 1 + \frac{2}{3} \cdot \mu_F + \frac{1}{6} \cdot \mu_W + \frac{1}{6} \cdot 0 \\ \mu_W = 1 + \frac{1}{5} \cdot \mu_F + \frac{1}{2} \cdot \mu_W + \frac{1}{5} \cdot 1 \end{cases}$$

$$\Rightarrow \boxed{\mu_F = 5.25.}$$

Problem 7 (12 pts. total)

A newscast covering the final baseball game between Sed Rox and Y Nakee becomes noisy at the crucial moment when the viewers are informed whether Y Nakee won the game.

Let a be the parameter describing the actual outcome: $a = 1$ if Y Nakee won, $a = -1$ otherwise. There were n viewers listening to the telecast. Let Y_i be the information received by viewer i ($1 \leq i \leq n$). Under the noisy telecast, $Y_i = a$ with probability p , and $Y_i = -a$ with probability $1 - p$. Assume that the random variables Y_i are independent of each other.

The viewers as a group come up with a joint estimator

$$Z_n = \begin{cases} 1 & \text{if } \sum_{i=1}^n Y_i \geq 0, \\ -1 & \text{otherwise.} \end{cases}$$

- (a) (6 points.) Find $\lim_{n \rightarrow \infty} \mathbf{P}(Z_n = a)$ assuming that $p > 0.5$ and $a = 1$. **Show your work.**

Solution: Note that

$$\lim_{n \rightarrow \infty} \mathbf{P}(Z_n = 1) = \lim_{n \rightarrow \infty} \mathbf{P}\left(\sum_{i=1}^n Y_i \geq 0\right) = \lim_{n \rightarrow \infty} \mathbf{P}\left(\frac{\sum_{i=1}^n Y_i}{n} \geq 0\right).$$

Since Y_i are i.i.d. with mean $\mathbf{E}[Y_i] = 2p - 1$ and finite variance $\text{var}(Y_i) = 1 - (2p - 1)^2$, one has, by Chebyshev inequality, for all $\epsilon > 0$

$$\lim_{n \rightarrow \infty} \mathbf{P}\left(\left|\frac{\sum_{i=1}^n Y_i}{n} - (2p - 1)\right| \geq \epsilon\right) = 0.$$

Take $\epsilon = p - \frac{1}{2}$, and the above equation implies $\lim_{n \rightarrow \infty} \mathbf{P}\left(\frac{\sum_{i=1}^n Y_i}{n} \leq (2p - 1)/2\right) = 0$. Therefore, $\boxed{\lim_{n \rightarrow \infty} \mathbf{P}(Z_n = 1) = 1}$.

- (b) (6 points.) Find $\lim_{n \rightarrow \infty} \mathbf{P}(Z_n = a)$, assuming that $p = 0.5$ and $a = 1$. **Show your work.**

Solution: Note that

$$\lim_{n \rightarrow \infty} \mathbf{P}(Z_n = 1) = \lim_{n \rightarrow \infty} \mathbf{P}\left(\sum_{i=1}^n Y_i \geq 0\right) = \lim_{n \rightarrow \infty} \mathbf{P}\left(\frac{\sum_{i=1}^n Y_i}{\sqrt{n}} \geq 0\right).$$

Since Y_i are i.i.d. with $\mathbf{E}[Y_i] = 0$ and $\text{var}(Y_i) = 1$, we can approximate $\frac{\sum_{i=1}^n Y_i}{\sqrt{n}}$ as a standard normal random variable when n goes to infinity. Thus, $\boxed{\lim_{n \rightarrow \infty} \mathbf{P}(Z_n = 1) = 1/2}$.

MIT OpenCourseWare
<http://ocw.mit.edu>

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.