Statistics Foundations: Understanding Probability and Distributions

by Dmitri Nesteruk

Course Overview

Course Overview

(Music) Hi everyone. My name is Dmitri Nesteruk, and welcome to my course on Statistics Foundations: Understanding Probability and Distributions. I am a quantitative analyst and software developer and have spent over a decade applying some of the skills presented in this course to analyze the financial markets, so I'm very happy to take you on this journey into the world of statistics. Now, we live in a world of big data, huge amounts of data generated by social networks and governments and consumers and markets, and all of this data needs someone to analyze it, which is why the profession of a data scientist becomes more and more popular nowadays. Someone needs to make sense of all of this data. In this course, we're going to cover all of the fundamentals that you need to understand in order to be able to efficiently analyze data, formulate hypotheses, and generally reason about what the ocean of data that's out there is actually telling you. Some of the major topics that we're going to cover in this course include the following. We going to talk about the notion of probability of course, we're going to take a look at random variables, we'll discuss statistical distributions, the idea of expectation, as well as things like covariance and correlation. By the end of this course, you'll be able to look at data and reason about it in terms of its descriptive statistics and possible distributions. Before beginning this

course, you should be familiar with mathematics at a school level, you should also be familiar with the R statistical environment. I hope you'll join me on this journey to learn statistics, with the Statistics Foundations course, here at Pluralsight.

Introducing the Concept of Probability

Course Introduction

Hello, and welcome to the course series on Statistics Foundations, and in this particular course we're going to study probability and various statistical distributions. My name is Dmitri. I'm a quant, and I'm going to be taking you on this journey to learn some of the fundamental statistics concepts. Now before we jump into the course itself, the first thing we have to discuss is what exactly is statistics and why should you care about it? Well, statistics is a branch of mathematics that deals with the collection, analysis, interpretation, presentation, and organization of data, so it's all about gathering data somewhere and then analyzing it and figuring out what the data actually means. Now the way that statistics typically works is something like the following. First of all, if you want to study some phenomenon you typically formulate a hypothesis. So for example, you might conjecture that smoking causes cancer, but of course to actually prove that smoking causes cancer you need some data, you need some hard evidence to present to people to show that smoking does in fact cause cancer, so what do you do? Well, you make observations. So one way or another you go out and you get information, you get the data regarding, for example, people's smoking habits and their medical history, as well as any factors which could influence the causality of your data, and so on and so forth. It's not a simple process, but once you get the data, what you do is you analyze the data and then you can make certain conclusions. So for example, you can accept or reject a particular hypothesis that you made, or for example you might realize that the experiment wasn't good enough and you have to go back and either modify the experiment, get more data, or perform some other adjustments to get better results. So here is what we're going to see in this course. First of all, we'll discuss the concept of probability, what probability is and how it works, and this is an introductory kind of module, which means in addition to probability itself we're going to be touching on a couple of other important topics. Then we're going to take a look at calculating the conditional probability of events, look at random variables and their various distributions, then we're going to discuss the concept of

expectation, and finally in the last module, we're going to look at some special statistical distributions. So let's talk about the structure of this entire course. So first of all, I have to reiterate that this course is a first course in a series of courses called Statistics Foundations, so it is the first steppingstone towards getting proficiency in statistics. Now the way the course is structured is as follows. In all of the examples I'm going to have certain theory lectures, we're going to have actual information presented on the slide where you get to learn about a particular concept, and in certain situations I'm also going to be doing live examples, and for these examples I'm going to be using the R statistical environment. The programming language that I'm going to be using doesn't really matter so much because R is just one of many choices that you can use for statistics, other choices include things like MATLAB and Julia and Python, and many other things. Now, one of the things that I enjoy doing, and you're going to see in this course, is a comparison between analytical results that we're going to derive or discuss as findings in the statistical space and comparing them to the results of simulations because certain claims that we would make, for example, about the random variables, it's nice to be able to verify them somehow, and simulation is one of the ways in which you can do that. It's a lot of fun to program simulations, and I'm going to have simulation examples throughout this entire course.

Module Overview

So, the goal of this module is for us to understand the concept of probability, but in addition to that, we also need to learn a few more basics from different branches of mathematics, and the branches we're going to cover are set theory and combinatorics, so we're going to have those and we're going to discuss probability using the apparatus that set theory and combinatorics provide us with. So here's what you're going to see in this particular module. First of all, we'll have a primer on set theory. Then we're going to discuss some of the terminology used in statistics, in particular we're going to discuss the idea of experiments and events. Then I'll introduce the concept of probability, we'll take a look at counting methods, these are useful for counting the so-called sample space, and we'll talk about what that is in a while, and we're also going to apply some of the ideas from combinatorics, which is another branch of math. And finally, we'll discuss the idea of probability of a union of events.

Introducing Sets

Alright, so we're going to begin our exploration of statistics with a dive into a different area of mathematics, an area called set theory. So, what is set theory all about? Well, set theory is a

branch of mathematics that deals with a mathematical construct called a set. Now, the kind of set theory we're going to cover is called naive set theory, and this is the kind of set theory which is explained using very simple terms without resorting to any kind of heavy math, any complex formula, nothing like that. We're going to have a very accessible kind of discussion on sets, and this is an alternative to a different kind of set theory, the so-called axiomatic set theory, so if you find that you enjoy set theory a bit too much, you can dive into axiomatic set theory, get yourself a book, and investigate that. But we're going to go with naive set theory because it's very easy to understand and digest. So, one question you might have is what exactly is a set? Now, a set is quite simply a well-defined collection of distinct objects, and by objects we mean virtually anything, it doesn't really matter what you have a set of. So for example here I have a set A, and this is a set of digits, we have digits 0, 1, 2, 3, and all the way up to 9. You can also have a set of letters containing letters a, b, c, and all the way up to z, or here for example I have a set S, which contains a bunch of strange looking mathematical operators. Now the way that sets are often illustrated is with the use of Venn diagrams, and that's certainly one of the ways that I'm going to be illustrating sets in this course. So let me show you what a Venn diagram looks like. So first of all, here I have a circle and a letter next to it, so this particular construct, the circle and the letter, is a representation of a set called A. Now in addition, you're also going to see some sets represented using a rectangle, that's just a different way of drawing a set. So here I have a set called S, and as you can see, the set A is inside the set S. So typically we have a set called S being a large rectangle representing all of the things which are possible, like for example, if you roll a die, the set S would represent all the possibilities, all the different die rolls that can actually happen. So, in addition to the representation of sets, of course, we have elements of a set. They're typically drawn directly on the set. So here I have x, y and z being members of set A, but because A is a subset of S, x, y and z are also members of set S as well, and this is how I'm going to illustrate elements of a set. Sometimes people put a point next to each of the elements of a set, but I'm not going to do this.

Set Membership, Null Set, Subsets

So, now that we know that x, y, and z are members of set A, we may as well formalize the relationship. So here on the left you can see I'm already using Venn diagrams, I have a set A and a bunch of numbers in it, 1, 2, 3, 4, and 5, and I also have the number 6, which is outside set A. So if we consider set A, and A consists of the numbers 1, 2, 3, 4, and 5, we can state that 1 is an element of A, and we have a special mathematical operator which looks like the letter E, which is used to represent the fact that 1 is in A, that's how you read it, left to right. And similarly we can say that 6

is not an element of A, I've drawn it outside the circle, so it's not an element of the set A. And once again we have a mathematical operator, which looks like an E that's been crossed out, and that represents the idea of something not being an element of some set. So here we're saying 6 is not an element of A. Now this notation is used in a large number of places, and one of the places where it's used, for example, is when you want to specify ranges of values or specify operations on ranges of values. So for example, if you take the following concept, if for example you want to state that you want to perform some operation on all even natural numbers, then the mathematical notation that you would use is something like this. So we have the upside-down letter A, which represents the idea of for all, so we have for all x, and then we have the conditions which are applied upon x, so in this case we have x being an element of the set of natural numbers, and the second precondition is that x is even. So, I don't have this notation in many places in the course, but this is something you'll see in math generally so this set membership symbol is going to appear all over the place if you go into mathematics. So one very special and unique set is the empty set, and that is quite simply a set which contains no elements whatsoever, and it's typically expressed as a crossed-out circle, and you can also define it as empty curly braces. So curly braces with no elements inside them. Now, an empty set is unique in the sense that there is only one empty set, and furthermore we can claim that for every element x, whatever x happens to be, x is definitely not an element of the empty set. Hopefully that's obvious because the empty set doesn't have any elements, so you cannot possibly be an element of a set which is empty. Now let's talk about this idea of subsets, and you'll notice I slightly changed the diagram here by putting the names of sets directly on the circle that represents the set. I've done this just to add additional clarity because sometimes it's difficult to figure out what's going on, and here we have a slightly more loaded, shall we say, diagram. So here on the diagram I have a set A, which is shown in blue, and it happens to be a subset of set B, and our definition of subset is that set A is a subset of set B if every element of A is inside B, so in terms of Venn diagrams you can consider a subset to be a circle which fits inside some other circle, for example. So, we can say that given that set A contains the elements 1, 2, and 3, and given that set B contains the elements 1, 2, 3, 4, and 5, yes, 1, 2, 3 are also elements of B because they reside on B as well as A, we can say that A is a subset of B. And once again, we have a mathematical operator which looks kind of like the E without the middle part, so to speak, so it's very similar to element of, but without that middle bar, and we can say that A is a subset of B. And similarly we have a set C, which contains elements 1 and 6, and even though the 1 element is also present in sets A and B as well, the element 6 is not present anywhere. So if C contains the elements 1 and 6, we can say that C is not a subset of B, and similarly we can say that B is not a subset of C, or A is not a subset of C, so all of these claims can be made. In addition, we have some fairly obvious definitions. So for every set

A, the empty set is a subset of A because if you look at our definition you can see that one set is a subset of another if all of its elements belong to another set and, of course, the empty set doesn't have any elements, so we automatically satisfy the condition of the empty set being a subset of anything. And any set is a subset of itself because once again we have specified that if every element of the set is part of the other set, then it's a subset. So A is, by definition, a subset of A.

Set Operations: Union, Intersect, Difference

Next up let's talk about the notion of a set union. So the union of two sets, A and B, is a set, so it's a brand new set, which contains all the elements from both A and B. So here on my diagram I have a set A containing numbers 1, 2, and 3, and I also have a set B containing 1, x, and an infinity symbol, or at least I'm going to call it that. And then the union of these sets is a brand new set containing the elements 1, 2, 3, x, and the infinity symbol. And you'll notice that even though the number 1 shows up in both sets A and B, it doesn't actually get repeated, so when you put something in a set, you're only keeping the unique values, you don't repeat the number 1 twice. Now it's also possible to define a union, which you can see has a kind of letter U here, it's also possible to define it as a large operator. So you can define a union of elements, A subscript i, where i goes from 1 to 3, and that would simply be a shorthand for A subscript 1, union A subscript 2, union A subscript 3, and we're going to be using this kind of notation in certain places. So in addition to the union, we also have the set intersection, which is highlighted in yellow on the diagram to the left, so that is essentially the intersection of the two geometric shapes. And formally defined, an intersection of two sets, A and B, is a set which contains elements which belong to both A and B. So only the elements which belong to both sets are in the intersection. So if A is equal to 1, 2, 3 and B is equal to 1, x, infinity, then A intersected with B is a set containing just one element, and that element is a 1 because that's the only element which is actually part of both sets. And we can also have it as a large operator once again. So you can have this very large cap, as we call it, so a very large cap with an index going from 1 to 3, and that is basically an intersection of every single element of the set A with a subscript going from 1 to 3, respectively. Next up we need to talk about set differences. Once again, given sets A and B, I have highlighted the set difference on the diagram on the left. So formally, given sets A and B, the set difference is the set of items in A, which are not present in B. So for example, if A is equal to 1, 2, 3, and B is equal to 3, 4, 5, then the set difference, which is sometimes expressed using a backslash, but there are alternative notations as well, the backslash is probably the most common, so the set difference between A and B are the elements 1 and 2 because those are the only elements which are present in A and not present in B. We can also provide some obvious

equalities, so the set difference between a set A and itself is the empty set because there are no elements which are both present and absent in your set. In addition, the difference between an empty set and some set A is an empty set, and the difference between A and an empty set is A because you're not removing anything by calculating a set difference on an empty set.

Cardinality and Set Complement

There are two more things that I wanted to mention, and that is cardinality and the set complement. So first of all, cardinality is a fancy word for saying how many elements there are in a particular set. So for example, if I have a set A containing elements a, b and c, then the cardinality of that set, which is represented by the sharp symbol, so a #A is equal to 3 because this set has 3 elements. In addition, we'll mention something called the set complement, and the set complement is quite simply a set of elements which are not contained in a particular set, but may be contained elsewhere. So for example, if you take the set E to have all the even natural numbers like 2, 4, 6 and all the way up to infinity, then the complement of that set, which is expressed as E with a superscript c, that's a small c by the way, would have the numbers 1, 3, 5, and so on, so that would be the set of odd natural numbers, and we can represent this as follows. So the complement of E is the set difference between the natural numbers and the set E. Now, you can, for example, consider a die roll. So, a die roll has 6 possibilities, 1, 2, 3, 4, 5, and 6, so if you take some arbitrary set A, which is a subset of the roll possibilities of a die roll, then if the overall set S has numbers from 1 to 6, then the set complement of A is the same as the set difference between the overall set of possibilities, 1 through 6, and the set A. So that would be the rolls of 3, 4, and 6.

Some Set Laws

I want to briefly mention some laws regarding sets. There are actually a lot more laws that you would encounter if you were doing set theory separately, but I wanted to mention some of them. So first of all we have De Morgan's laws. So, that basically says that the complement of a union is the intersection of complements, and the reverse goes for the intersection and the union, respectively. We also have distributive properties, kind of like, you know, with multiplication you have A times B plus C equals AB plus AC, well it's kind of similar here in the sense that A intersected with the union of B and C is the same as A intersected with B union A intersected C, so these are useful distributive properties. And we also have some additional properties as well, so given that you have two sets, A and B, you have the intersection of A and B and the

intersection of A and the complement of B are disjoint, and I guess at this point I should mention what disjoint is. Disjoint means the sets do not overlap. Another definition that might come in useful is that A can be defined as the intersection of A and B unioned with the intersection of A and the complement of B. Once again, this is something that can come in handy. Now, as I mentioned, if you go and get yourself a set theory book or just read an algebra book, which is more sort of towards in-depth math, shall we say, then you're going to see a lot more laws and a lot more different kind of set and subset operators as well, so if you're interested in these things, consult a good textbook, a good mathematics textbook on the subject.

Experiments and Events

We're now going to discuss the idea of experiments and events. So we'll begin with events. So if you consider an experiment where you roll a balanced, 6-sided die, then you don't really know what value you're going to get, you might get a 1, you might get a 6, you might get some of the other values. So what we do is we call this value that you get a random variable. Now there are different kinds of observations you can make as you roll your die. For example, you can observe that a player rolled a 6, and this is something that we call a simple event. Now, the reason why it's called a simple event is because you cannot decompose the player rolling a 6 into a further combination of events, it's kind of singular, impossible to separate kind of event. But, it is possible to get something more complicated. So for example, you might observe that a player rolled an even number. Now, what this means in practice is that the player either rolled a 2, a 4, or a 6 because these are the only even options if you have a 6-sided die, and this is something that we call a complex event because this is the kind of event that can be decomposed into simpler events, in this case it can be decomposed into a union of events, 2, 4, and 6, corresponding to the die rolls that might have happened. Complex event processing is actually one of the branches of computer science, and we have special packages and tools for exactly this purpose, but it doesn't really affect the discussions in this particular course.

Sample Spaces and Points

Another very important term that we need to discuss is sample space, and the sample space is quite simply a set of all the possible values of a random variable. So for example, for a die roll, you might describe the sample space as a set containing events E1, E2, E3 and all the way up to E6 where E with subscript n corresponds to an event where the n value has actually been rolled on the dice. Now, each of these elements of the sample space is also typically called a sample point.

Set Operations on Events

So now that we've met sets, we may as well take a look at how to perform set operations on events, and for this we're going to consider a die roll. So here I'm modeling a die roll with six possible outcomes, and I have my Venn diagram on the left with these six sample points. Now, let's suppose that I observe an even die roll, so that constrains the set of sample points to E2, E4 and E6 as highlighted in green on the left because these are the only die rolls where the number is actually even. Now in addition, I can have another constraint, like for example if I observe a die roll that is greater than 3, then obviously it has to be a 4, a 5 or a 6, and that is shown as the blue triangle on the left. And now the interesting part, the use of set theory to actually kind of constrain our search space even more. So for example, let's suppose that I observe a die roll which is even or greater than 3. So in this case, what happens is we take the union of the set of even numbers, which is green, and the set of numbers greater than 3, which is blue, and so we get the values E2, E4, E5, and E6 because that is the union, and that is the area which is encompassed by both the triangle and the rectangle on the left. And similarly we can, for example, observe a roll that is both even and greater than 3, and in this case we need an intersection. So we intersect the set of even numbers, and we get just 2 of them, we get E4 and E6 because both of them are even and both of them are greater than 3. Finally, we can observe an impossibility. We can observe a roll that is odd and greater than 5. Now the problem with this is that the set of numbers which is odd, which is shown in blue, and the set of numbers which is greater than 5, and that's only one number, E6, they don't actually intersect. So if you take the intersection of two disjoint sets, then what do you get? Well, you get an empty set, and this means that there is no such event and that this observation is actually impossible because no number, when rolled on a die, is both odd and greater than 5.

Independence of Events

Another thing worth discussing is this idea of dependence or independence of events. So if two events that are being observed, one after another, don't influence one another in any way, they are called independent. So a typical example is something like a coin flip or a die roll. You flip a coin once, you get a result, you flip a coin again, you get a different result, but the previous result doesn't affect the subsequent result, they're completely independent. And the same goes for a die. So you roll a die, it gives you a 6, you roll the die again, it gives you a 1, but its previous result doesn't affect the subsequent result. Now, this isn't always the case because sometimes subsequent events are influenced by the previous events, and in this case, we call these events dependent. So for example, suppose that you have a bag with a bunch of colored balls in it. Now

let's suppose that you pull out one of the colored balls. The subsequent events, the subsequent attempt to pull out a ball, is affected by this attempt that you made just now because you removed one of the balls so effectively you've reduced the sample space of the experiment. So as you reduce it, you basically change the probabilities, you change the way that the experiment is affected, so in this case we call these events dependent.

Demo: Set Operations

Let's take a look at how you can use the R statistical environment to perform some set operations. Now, this isn't something you would be doing much in real life, but since we're studying sets, why not calculate a few things? So I'm going to have two vectors. First of all I'll have a vector from 1 to 4, and then I'll have another vector going from 3 to 6. So it's going to be x and y, respectively. So here we go, 3 to 6. And now, I can treat both of these vectors as sets and I can perform certain set operations and we're going to begin with set can be set is equal to another set. Now this is done using a function called setequal, you can see that these functions are available without importing any packages, so all the set operations are available right out of the box. We can compare the set x with itself, and if I do that obviously I get the value of true, as you can see. I can compare set x with set y, so we can do x comma y here, and when I execute this, predictably enough, I get a false. But here's something interesting. If I do setequal and I compare set y with another set where the numbers from 3 to 6 are written in reverse order, so I have 6, 5, 4, and 3, like so, I'm going to get a true, and this illustrates very well that sets don't really care about the order of the elements in the sense that both of these sets, even though the ordering of the elements is different, in these 2 sets they are considered identical from the point of view of set theory. Now, we can test whether a particular element is inside the set, so for example, we can say is. element and we can check, in fact, whether the value 2 is in x, and if I execute this I get the value of true. And similarly, I can also use the %in% operator to do exactly the same thing. So I can do 2 %in% y, for example, and I'll get the value of false because 2 is not an element of y. Okay. Another thing that we can do is we can check whether one element is a member of some set or other, we can do it on an entire vector, so what I can essentially write is something like is. element x, y. Now what this does is it takes every element of x and it checks whether this particular element is in y, and as a result it doesn't just give you a single true or false value, but it gives you a whole array of values. So as I execute this, I get false, false, true, true because the first 2 elements, 1 and 2, are not present in y, but the elements 3 and 4 are in fact present in y, and so I get the true values here. Now the common set operations that we looked at such as union, intersection, and difference are fairly easy to do so. To do a union you just type

union x, y and in this case we get 1, 2, 3, 4, 5, 6, all the elements from x and y. If I want to do an intersection, I can type intersect, and once again have x, y and execute, and I get 3 and 4 because that's the only 2 elements which are present in both x and y. And similarly for the set difference, it's setdiff, and once again I can do x and y, and in this case I get 1 and 2 because these are the elements which are present in x, but not present in y. Now, what you can do with set operations is you can mix vector modes, and let me show you how this works. So I'm going to make another array, so I'll have a set of values, and this time I'll have letters, I'll have a, b and c. So I'll initialize this, and then what I can do is I can, first of all, show you the mode of x, which is numeric, and I can also show you the mode of z, which is, of course, character. Now what I can do is I can do a union, so I can for example say that a is a union of x and z, and in this case, this operation succeeds. If I look at a, you can see that it's a collection of all the elements so 1, 2, 3, 4, a, b, c, and if I do mode of a, you'll see that this is a mode of character. And you can do similar things, like for example intersect, you can intersect x and z in this case, and this will give you 0 characters, so it still converts them to characters, but of course there is no common element between a set of numbers and a set of letters, so you get an empty character array in this case. And you can do a union of a union. If you want to union three things together for example, I can do union of a union of x and y and z as well, and if I execute this, as you can see, I get 1, 2, 3, 4, 5, 6, a, b and c.

Introduction to Probability

The key topic of this particular module, at least, is probability, so that's what we're going to discuss now. Now probability is quite simply how likely a particular event is to occur. So for example, if I roll a 6-sided die, there is no way I'm going to get an 8, am I, because it's a 6-sided die, but there is some likelihood that I will get a 1 or a 6. So let's try to actually formalize this definition of probability. First of all, I do it using plain English and then we'll go more formal and try to do it using mathematical terms. So, probability of an event is a number, so it's a number which is somewhere between 0 and 1. So this kind of constrains the range, probability cannot be negative, it cannot be greater than 1. Now the value of 0 in probability corresponds to something that is extremely unlikely and almost never to occur. So for example, let's suppose you flip a coin 1, 000 times, so what's the probability that you get 1, 000 heads? Well, it might not be exactly 0, but it's sufficiently low so as to claim that it will almost never happen. And similarly we have the value of 1. So if you flip a coin, for example, what's the probability that you will get a 0 or a 1? Well, the value is a 1 because it's almost certain, it's almost surely that if you flip a coin it's going to land on the flat side as opposed to its edge, for example. So typically, the way we record probability is using the capital letter P. So we say P, then round bracket, the event that you're trying to assess,

equals value. Sometimes though we express probability not using a number from 0 to 1, but using percentages where that number is effectively multiplied by 100. So for example, when somebody tells you that there is a 50% chance that something might happen, that is the probability of 0. 1 or 1/2, and similarly if somebody tells you they're 90% certain that something happens, the probability value is 0. 9, or 9/10. And this is used in many places. So for example in Microsoft Excel, if you write 10%, for example, that would be changed to 0. 1 behind the scenes. And incidentally, in Excel it doesn't really have to mean probability, it can be part of some sort of financial calculation, for example, but still, that's how the conversion works between percentages and ordinary values. So in terms of probability, you take the percentage value and simply divide it by 100.

Rules of Probability

Now here are some of the more formal rules of probability. So first of all, the probability of any event A is greater than or equal to 0, so that's rule number 1. The second rule is that the probability of events in the sample space is 1, which is recorded as P of S equals 1. Now, this probably needs some explanation. So suppose you flip a coin, for example. The probability of a coin actually giving you some sort of value is 1 because it's almost certain that if you flip a coin you're going to make an observation, and that observation will be meaningful, and S here represents the set of all the different outcomes. So the probability of one of these outcomes happening is equal to 1. Now the third rule is a bit more complicated, and it states that if you have a sequence of events, A1, A2, and so on, and they're mutually exclusive, which means that they don't overlap if you present them as sets, then the probability of a union of these events is equal to the sum of their respective probabilities. Now we'll come back to this rule a bit later, but don't worry about it for now.

Probability Examples

So let's take a look at some examples of probability. So for example, let's consider our favorite coin flip where we have two possible events, so you either get a head or a tail. Now by definition, the probability of the actual sample space is equal to 1, but in addition, if it's a fair coin, we know that the probability of a head is equal to the probability of a tail because both of these events are equally likely. So what we end up with is we end up with a system of equations. So the first equation arises from the fact that the probability of a union is equal to the sum of probabilities. Now we know that the union here, which is represented by the letter S, is the union of the head

and tail events, so their sum is equal to 1. Now, also we know that since probability of heads is equal to probability of tails, probability of heads minus probability of tails is equal to 0, and we now have a 2 x 2 system of equations, which gives us the probability value for rolling either a head or a tail, and in both of these cases the probability is actually one-half. Now, we can also make other statements about the probabilities related to coin flipping. So for example, the probability of flipping both a head and a tail in a single flip is actually equal to 0, because the coin cannot land on both of its sides at the same time, and once again the probability of getting either a head or a tail is quite simply the probability of the sample space because these are the two options which the coin affords us, and by our definition of probability, this value is equal to 1, because we know that the coin will definitely land on either the head or the tail.

Demo: Basic Probability

Like I said at the beginning of the course, we're going to do simulations in order to test some of our assumptions and results, and in this particular simulation, and this is going to be a simple one, we're going to see whether or not the probability of a coin toss being a head or a tail is in fact 0. 5. So I'm going to do two implementations of the simulation. The first implementation is going to use the random number generator that's built into R, and the second is going to do pretty much the same thing, but we're going to use syntax which is much closer to what statistics is like. So first of all we need to decide how many coin tosses we're going to simulate, so that's going to be the number N, and I'll start at 100. So we're going to simulate 100 coin tosses, we're going to sum up the results, and we hope that half of those, or roughly half of those, will be heads or tails. Remember, half is 0. 5, that's what we expect from theory. So, let's do this using the random number generator first of all. So I'll have a variable called flips, and I'm going to assign it to the following. So I'm going to use the uniform random number generator, and here I'm going to specify the number of flips, in this case is going to be N, that's 100, and then we'll keep the minimum as 0 and the maximum as 1, I'll put those here explicitly so that you can see what's going on. So we're going to generate 100 random numbers from 0 to 1. And to turn them into Booleans I'm going to compare them with 0. 5, that way we're going to get a bunch of true and false values depending on whether each number is greater than 0. 5 or not. So if you get a 0. 7, that's going to turn into a true, if you get 0. 1, that's going to turn into a false, so flips is going to be an array of Boolean values. And what we expect is we expect that if we sum all the elements here, if we sum all the elements in flips, now remember, true becomes a 1 and 0 becomes false, or false becomes 0 rather, if we sum up all of these elements then we expect the sum of these elements divided by the number of elements to be roughly 0. 5, more or less. So let's actually

execute all of this, and as you can see, we're getting 0. 45, which is not exactly what you would expect, because in a perfect scenario where the probability is exactly one-half, you would get 50 heads and 50 tails. Of course this is a simulation, which is why you'd never get a perfect result, but you can change it by increasing the number of coin flips, for example. And this time around, if I execute it, I get 0. 482, and you can experiment with this number to see what kind of result you're getting. So this is how you would simulate the process using just ordinary uniform random number generation. Now we haven't really talked about the concept of a random number in any great depth, just assume for a moment that runif generates a number from 0 to 1 and it generates a whole array of them, depending on how many elements you want. The uniform part means that all of the elements have equal probability. Now we're going to simulate the same thing, but we're going to do it using language, which is more consistent with what we've been discussing throughout this entire course, because we're going to define a sample space. So to define a sample space for this experiment, I'm going to say sample. space, I'm going to introduce a variable called space in this case, and I'm going to say that the sample space consists of 0 and 1. So these are codified values, so 0's going to be maybe heads and 1 is going to be tails. Now if I want to perform sampling from the sample space, I'm going to use the sample function. So this time round I'm going to say flips, and I'm going to assign the following. So I'm going to use a function called sample, and this performs sampling from the sample space. So it takes a value at random from the sample space, and you take as many values as you need for your particular simulation. So here we specify the sample space that we're actually going to be using, we specify the number of elements, the size, then we specify whether or not we perform replacement. Now remember, we're tossing the same coin again and again and we replace it at each turn. So here we say replace = TRUE, and in addition we can specify the probabilities, the likelihood that you will have a 0 or a 1. So in this case what I can do is I can say prob =, and I can specify that both of the probabilities are equal to 0. 5. There we go. So having specified this, I perform the same calculation as I did above, I sum all the flips and I divide them by the total number. So let's take all of this code, execute it, and as you can see, we're getting 0. 494, which is also pretty good. So, this is how you can perform a simulation of generating random coin flips and then looking at the results, and this way we can see that the results we're getting agrees with the theory value of one-half that we were expecting.

Discrete and Continuous Probability

Another thing worth mentioning is the difference between discrete and continuous probabilities. So far we've been looking at discrete cases, and those are cases where you have a finite set of

unique values that can be considered. So for example, in the case of a coin toss, you can get a head or a tail, that's it, these are the only two values. In the case of a die roll, you typically have six values. In the case of, let's say, the number of cars that a typical household has, it might be 0, 1, 2, 3, and so on, so it's a finite number, we don't go to infinity, we don't go to any kind of infinitesimally small values, it doesn't get fractions, like a family cannot just have 1. 5 cars for example. Well, I suppose you could extend reality to allow car sharing and whatnot, but in the general case that's a set of discrete values. And for all of those values we have a very simple formula that the probability of an event, the probability of, let's say, a die roll being the number 3, is 1 over N given that all of the N possibilities are equally likely. So if you have a fair, balanced die that you are not cheating with, then the probability of rolling a 3 or a 5 or a 6 is 1/6 because N is equal to 6. Now in addition to these discrete cases we also have the continuous case where the set of possible values of something is infinite, and a typical example is a person's height. If you measure a person's height you can never measure it exactly, first of all, because it might be a very, very large number in terms of the actual number of decimal places. We can certainly round it to some particular value. But if you take, for example, the amount of rainfall in a given day, that can once again be a value, the probability of rain, for example, being equal to exactly 5. 0 is 0 because you cannot measure right down to the molecule or subatomic particle to make sure that that's exactly the value, so in this case we can generalize. And in the same vein, the probability that you're exactly 1. 77 meters tall, right down to the kind of molecular level, is equal to 0. So in the continuous probability scale, the probability of any distinct value is equal to 0 because the number of different values is considered to be infinite. So because it's infinite, you are dividing by infinity and division by infinity gets you 0 in terms of probability. So if you're talking about people's height, for example, it might be simpler to either round people's heights, so round them to the nearest millimeter, for example, or for example you can talk about intervals. So you can, for example, measure the number of people whose height is somewhere between 1. 77 and 1. 78, that way you don't have to be so precise in terms of measuring their actual height.

Counting Sample Points

In order to be able to say anything about the probability of a particular event, which is part of some sample space, you have to be able to tell how big the sample space actually is, so we're concerned with how to count the number of sample points that exist. Now coming back to the very simple coin flip here, we certainly know that the number of elements in the sample space here is equal to 2, you either roll a head, or throw a head, or you get a tail. Now let's suppose that we complicate this example. So we decide to flip a coin three times, or we decide to take three

coins and flip them once, which is exactly the same thing. And this is something that's typically called sampling with replacement. And the reason why we're calling it with replacement is after we flip the coin, we kind of replace the coin and use it once again, so we don't suddenly get rid of the coin, we don't take a different coin, and this is important, this with replacement and without replacement is going to be a continuing theme throughout our entire discussion of probability. Now, when we flip these three coins the question is, how many different arrangements is it possible to actually get? Now, the simplest way is to just count, so it's a very direct approach. You go ahead, you list all the different arrangements of the three coin flips, and you just count them. So here they are, head, head, head, head, head, tail, and so on. As you can see, we can now see the overall sample space of this experiment. So in this case we have eight different points, I believe, in this experiment, and we can start asking questions about this experiment. So for example, I can ask a question, well, how many arrangements of these three coins begin or end with a head? So, if we use the mathematical notation, we want the cardinality of a set of rolls where head is first unioned with the set where the head is last. Once again, if we're just doing the counting, then we can simply count them, we can count them as shown here, and as you can see, there is precisely six of them. So provided you can easily list all the sample points in your experiment, provided you can list all of the sample space, in fact, you can just pick out and count the number of points. Unfortunately, sometimes it's not so easy, typically because the number of sample points is actually very large. Now, this example can also be visualized, we can visualize this example using a tree that you can see here on the right. So it's a tree where essentially each branch represents a possible event, so as you can see, as we start on the left-hand side, the branches represent heads and tails and it kind of goes on for however many coin tosses you actually do. And then the final branches, kind of the final terminal values, represent all the possible outcomes, and that is exactly what we had listed on the previous slide. And clearly as you can see from this arrangement, given that you are doing N different tosses, the total number of outcomes of this experiment is 2 to the power of N because it's a binary tree and the number of leaves of a binary tree of height N is 2 to the power of N.

Multiplication Rule

Okay, so maybe counting isn't so exciting, let's see if we can actually start using some math to figure out how many events we actually have in the sample space. So here is another example. Let's suppose that we need to travel from point A to point C and we need to go through point B. Now, there are three ways of getting from point A to point B, which are illustrated using the orange arrows, and we can also illustrate them using a set notation as well, and there are two

ways of getting from B to C as well, which we can also illustrate using a particular set. So the question then is, how many ways are there in total? Well, what we can do is we can certainly do a table, we can just draw a table which lists every single possible path, and as you can see, given that this table has 3 column headings and 2 row headings, 3 times 2 is equal to 6, and expressed using, once again, cardinality notation, we can say that the number of different options for traveling from A to C is equal to the cardinality of set AB multiplied by the cardinality of set BC, which is equal to 3 times 2 which is equal to 6.

Permutations

Alright, so far all of the examples have been rather simple, so how about looking at something more complicated? So here I have a situation where I have a bag, and this bag has three different colored balls, it has a red ball, a green ball, and a blue ball, and we draw one of these balls after another, and we draw them without replacement. So after you draw a ball from the bag, you don't put the ball back in before you actually draw from the bag again. So, you actually get to keep the ball, and the question is, how many possible draws are there as you draw the balls from the bag? Well, let's think about it. So, let's suppose that you have a bag, it has three different colored balls in it and it's your first draw. So on the first draw, there are three different options, you either draw the red one, the green one, and the blue one, and whichever ball you draw you actually keep, so you don't put it back. On the second draw, you've already taken one ball out, so you've restricted yourselves to just two possible options. So on the first draw we have three options, on the second draw we have two options, and on the last draw we don't really have any choice, there is only one option because you've taken two balls out, so only one ball is left in there so there is less variability here. Now, if we multiply these values using the multiplication rule, we're going to get the following. We're going to get the total number of possibilities being equal to 3 times 2 times 1, which is equal to 6. And by the way, here are the actual possibilities in this case. Now, the official mathematical term for what we're calculating, the number of different arrangements, is called permutations. So given a bunch of elements, all of the distinct arrangements of these elements are called the permutations, so we talk about the permutations of a set. Now the number of ways in which you can arrange N elements out of N possibilities, so for example, you have a bag with five different colored balls in it and you pull out exactly five, one after another, is expressed mathematically as follows. So it's the capital letter P where the subscript is n, n, and I'll explain why that is in a while, but essentially it's a product of all the numbers starting at N and ending at 1, which is also represented as N factorial. So that N! next to it is called N factorial, and it's quite simply a product of all the numbers from N down to 1. Now of

course there are some assumptions here, like for example we assume that N is a non-negative integer, and we also, by definition, assume that 0 factorial is equal to 1; this can show up in certain places. Okay, so in this particular example we might be drawing, let's say, five elements out of five, but let's try to generalize it a bit more. So let's suppose that you have five balls, red, green, blue, white, and orange, but you only draw three balls out of a hat or a bag without replacement, another question is, well, how many arrangements are there in this particular case? So once again, we can start thinking about it using the multiplication rule first of all. So forget the factorials for a second, let's think about the multiplication rule. So in the first draw, you have five possibilities. There are five balls, and you draw one of them, so five possibilities on the first turn, but as you've taken out a ball, then there is four possibilities, so that's 5 times 4. And on the third turn there are three possibilities. So the overall number of different options is 5 times 4 times 3, which is equal to 60. Now, 5 times 4 times 3 doesn't look like a factorial, does it? And what we really want is we want some sort of formula for taking these values n and k where n corresponds to the overall set of sample points and k corresponds to the ones that you're actually kind of pulling out and generating, so to speak, we want some sort of formula, and we can actually get this formula. Here is a neat trick that you can do. So 5 times 4 times 3 can be also multiplied by 2 by 1 on both the nominator and denominator, so I'm cheating a bit, I'm multiplying these numbers by 2 by 1 so that the top part becomes a factorial, and I'm dividing them by 2 by 1, and guess what, 2 by 1 is also, it's a factorial, it's a factorial of 2, in actual fact. And if we now go from numbers to n and k, we get this nice formula. So we have the number of permutations of k elements out of n, which is expressed as P underscore n, k, and that's calculated as n factorial divided by n-k factorial.

Permutation Examples

So, let's take a look at some examples of permutations. So for example, let's suppose I have a set of different letters, a, b, c, d, e, and I want to find out the number of unique five-letter words which can be arranged from these letters. Now obviously if I'm allowed to repeat the letters, which means that we are doing replacement, so we can pull out a letter, but then we can replace the letter, then the total number of possibilities is 5 to the power of 5 because you can have an element in each of the 5 positions. So that's 5 options in the first place, 5 options in the second place, and all the way to the last, so that's 5 to the 5. But if you're allowed to use each letter exactly once, then the number of possibilities reduces to 5 factorial, which is equal to 120. And then of course, if you decide that instead of five-letter words, you want to create, let's say, three-letter words, then you use that P formula which we've just looked at. So that's P 5, 3, and that's 5 factorial divided by 5 - 3 factorial, which is equal to 60. So, one very famous problem which is

related to permutations is called the birthday problem. So, let's suppose that you're in a room with k people, let's ignore yourself, let's imagine you're looking at a room with k people. Now the question is, what is the probability that at least two people in the room have the same birthday? Now, we're going to do a couple of simplifications here, we're going to ignore leap years and people being born on February 29th and whatnot, and we're also going to ignore any kind of seasonal variations because in actual fact, people don't get born uniformly throughout the year, some couples in fact really want their children to be born on summer months, for example. So we're going to assume that everything is equally probable, shall we say. Now, obviously if you have the number of people more than 365, and that has to be a rather large room, then the probability becomes a 1 because it's impossible to have all of these people born on different days, by definition. So we're going to assume that the number of people is somewhere less than or equal to 365. Now, we know that the number of arrangements when you actually take k people out of this room, the number of arrangements is using the P formula, so it's the number of permutations of k elements out of 365. And we also know the total number of arrangements that you can actually have, that is, 365 to the power of k. So the probability of no 2 people having the same birthday in this room is that P value, so P 365, k, that's essentially the permutation calculation, divided by 365 to the power of k. And then of course if you want to calculate the probability of two people having the same birthday in this room, then all you have to do is take this calculation, take one and subtract the result of this calculation from it. So the probability of 2 or more people having the same birthday, having a birthday on the same day in this room is 1 minus the set of permutations divided by 365 to the power of k. This is actually an interesting discussion. So, let's take a look at how this is done in R, and we're actually going to do a plot and see where this leads us.

Demo: Birthday Problem

What I want to do now is I want to show you the probabilities related to the likelihood of finding two people with the same birthday in a room full of n people, and n is something that we're going to vary. Now, there is a bit of a problem in using R for these calculations straightaway, because if you remember, the calculation of both permutations, as well as 365 to the power of k, these are huge numbers, and unfortunately the default data types in R simply cannot handle the range. So we're going to use an external library, which will allow us to specify the precision of our data types, and thereby we'll use data types which are large enough to perform these calculations. So here I'm going to add a require statement and we're going to use a package called Rmpfr. So that's the package which actually allows you to define these very large numbers. Now let's define

the function for calculating the permutations. So it's going to be a function which takes n and k, and what we're going to do is we're going to divide the factorial of n by the factorial of n - k, except that, as I just said, we're going to use these rather large types, and as a consequence, we're going to use a different function. So instead of just using the factorial function, we're going to be using the factorialMpfr function. So we're going to return factorialMpfr, and notice I'm not getting the completion here because I haven't loaded the actual library yet, so let's do that first of all. And now that I have, hopefully I will now get the completion here, Mpfr, so I'm going to provide n. So n factorial is divided by factorialMpfr of n - k. So that's how we do the permutation function using this expanded precision. Now, I'm going to define the count, so the number of elements that we're going to process, from 1 to 50, so the limit is going to be 50. I'm also going to set up a double array with just 50 elements of type double that we're going to fill in later on, so that's going to be called p, that's going to be the probabilities, so I'll say that's a double with a size of count, and then I'll just to do a loop to perform the actual calculation. So for k in 1:count. So that's where we're going to actually calculate both the permutations, as well as everything else. So first of all, what I'm going to do is I'll have a temporary variable called d, which is going to be using the multiple precision definition, so it's going to be a very large number, but then I'll collapse it to a double because as you can see I've defined the array of probabilities to be of type double so that we can plot it, because plotting these arbitrary precision huge numbers is rather difficult, it's too difficult for us. And in any case, those numbers aren't going to be large, because essentially we have a huge factorial here, but it's divided by an equally large factorial here, so there is no problem. Okay, so how do we calculate d? Well, if you remember, it's 1 - permutations divided by 365 to the power of k, so let's calculate exactly that. So it's 1 - permutations, from 365 elements, we take k elements, and then we divide it by 365 to the power of k. Now here is where the problem lies. So 365 to the power of anything is a huge number, so we're going to be using that Mpfr library and I'm going to define 365 as an Mpfr number. So mpfr, mpfr 365, and here I have to specify the number of bits I'm going to allocate to this number. I don't know many bits to specify, but I'm going to specify 1, 024 because that seems like a suitably large number to define the value of 365 to the power of 50, for example, that's going to be the largest value. So we calculate this value to the power of k, like so. And in addition what I'm going to do is I'm not going to keep just raw probabilities, I'll multiply by 100 so I get the percentage values. It's just a little bit easier for us to on this tab. Now what I'm going to do is I'm going to initialize the probabilities array at point k, and I'm going to convert from this Mpfr number, this huge number, I'm going to treat it as an ordinary numeric. So this is done using asNumeric, so just provide the number d here, and now that we have an array of double values in p, we can actually plot this. So I'm going to plot, so I'll plot the x axis from 1 all the way up to count, the y axis will have the

probability values, the x label is going to have number of people in the room, and the y label is going to have probability in percentages. Okay, so let's select all of this and execute this, and hopefully we get a nice graph. Here it is, let me just drag this over so we can see it better. Okay, so as you can see, it's not a linear relationship, it's a bit curvy, but the consequences, once you get to about 50 people in the room, it's almost certain that 2 of them have the same birthday. If you want, you can expand this range a little bit going from 0 to 100 people, for example, and in this case you will see the graph flatten out. So at around 60 people it just goes to a virtual certainty, so this is the result that you pretty much would expect as you get more and more people. But maybe it's a bit surprising to you that the certainty arises at around 60 people and not at 300 people, or something like that. But that's the value that we get using probability theory, and it's a good illustration of how to perform actual calculations and visualizations in R as well.

Combinatorial Methods

Alright, so now we're going to talk about combinatorics, and combinatorics is yet another branch of mathematics which is concerned with the study of finite or countable discrete structures. I know this doesn't say much, but in our case the reason why we want combinatorics is to learn how to measure the size of the sample space, that's what we've been talking about throughout the past few examples, so let's continue with this idea. So, let's suppose that you have, once again, a bag of colored balls, you have red, green, blue, white, and orange. Now, if we pull out two balls from this bag after replacement, the number of different arrangements is 5 factorial divided by 5 - 2 factorial, which is 5 factorial divided by 3 factorial, which is equal to 20. So we've talked about permutations and we know that this is the number of arrangements that we can have. However, one thing that we're neglecting and we're going to kind of reintroduce now is this idea that we might not care about the order in which the elements were actually pulled out. So when you pull out a red ball followed by a blue ball, let's suppose that we decide that we consider this case to be equivalent to the case where we pull out the blue ball followed by the red ball, and we want both of these to count as a single pick, as opposed to two picks. And the question is, well, how can we now calculate the number of unique picks of balls out of this particular bag? This is a bit of a problem because we need to, what we need to do effectively is we need to find the number of subsets of a particular set, because remember, in the context of sets, in the context of a particular set, the order doesn't really matter. The set which contains numbers 1, 2, and 3 is the same as the set which contains numbers 3, 2, and 1, so we want find the number of subsets of a set. Okay, so the rules of the game are as follows. We are given a bag of balls where we have R, G, B, W, and O, red, green, blue, white, and orange balls. Now, we can calculate the total number of

picks from this bag, let's suppose that we're taking out just two balls, so we can use the permutation formula to figure out what the actual set of possible picks, including those interchangeable ones are. So, that is step number 1. Now step number 2 is once you have a pick, we can also calculate the number of different rearrangements within this pick, and since we're picking 2 balls, we can then have 2 factorial possible rearrangements within that group that we've picked out. So what we can do is we can reduce the total number of picks by the number of rearrangements of a particular pick, so essentially we take the permutation calculation and we divide it by k factorial, we divide it by how many times you can actually rearrange within that group that you've picked out. So in our case, we choose 2 from 5, that's the capital letter C, we have the capital letter P for permutations, the capital letter C stands for combinations, or sometimes it's called choose, so we choose 2 elements out of 5, and that is calculated using the formula that you see here on the screen. So we take the number of elements as a factorial, we divide it by n - k, and that n factorial divided by n - k is the permutation formula, but we also divide it by k factorial, so applied with the numbers 5 and 2, that gives us 5 factorial divided by 2 factorial, 5 - 2 factorial, which is equal to 10. So what this means is that if you have a bag of five balls of these colors, there are 10 possible different arrangements of 2 balls which are taken out from the bag, and here they are. Here is a listing of all the possible combinations that you can actually pull out of a bag.

Binomial Coefficients

Now, this C notation that we've used, C n, k is also written using round brackets with n on top and k down below, and these are so-called binomial coefficients. Now, the reason why they're called binomial is because of the binomial theorem, and that's basically a theorem which states that if you have 2 values, let's say x plus y, taken to the power of n, that is equal to this particular sum going from 0 to n, and it's a multiplication which we don't really care about so much, but what we do care about is the values of n and k, which show up as a coefficient. So effectively this calculation of the number of combinations actually shows up in the different kind of algebraic setting because if you want to take a sum of numbers to some power, then you can do it as a sum where this n over k is actually the combination calculation that we looked about previously. Now, for all values of n, n over 0 is equal to n over n is equal to 1, and we can also state that n over k is the same as n over n - k. These are just small observations on the calculation which might come in handy. And the definition of n choose k and the definition of n over k are exactly the same, it's the same calculation just using different notation. We're actually going to be using the binomial coefficient notation rather than the capital letter C for most of our examples, so let's take a look

at a few examples where this kind of calculation actually shows up. So let's suppose you have a coin and you toss the coin 10 times and you record the result. Now the question is, what is the probability of getting exactly four heads? Now we know that since you flip it 10 times there are a total of 2 to the power of 10 arrangements, which is a rather large number, but we also know that each arrangement is actually a choice where you have the 4 heads and you somehow distribute them among the 10 different tosses. So overall, if you had to distribute 4 heads among the different tosses, you would have 10 over 4 different arrangements. So the probability of tossing a coin and getting exactly four heads is the probability of that pick, that pick of four heads divided by the overall number of different cases. So that's 10 over 4 divided by 2 to the power of 10, which is equal to roughly 0. 2. Now, in addition to calculating the probability of exactly four heads, we can also calculate the probability of having four heads or less. And in this case what we would do is we would simply calculate the probability of 4 heads plus the probability of 3 heads plus the probability of 2 heads plus the probability of 1 head out of the 10 coins plus the probability of 0 heads, and that is a fairly obvious calculation. So in this case you would sum them up and you would get a slightly higher value, 0. 377 in this case. Okay, here's another, more complicated example. Consider a class of students, you have 15 girls and 30 boys. Now let's decide that you decide to pick 10 children at random. Now what's the probability that you're going to pick exactly 3 girls in that 10-person group? Now the number of ways in which you can pick 3 girls just from the girls is 15 over 3, because there are 15 girls and you're picking 3 of them. Now, the number of ways of actually picking 7 boys from the 30 boys that you have, and the reason why we have the number 7 is because if there is 3 girls, there is obviously 7 boys, so the number of ways of picking 7 boys from 30 boys is 30 over 7, and so the overall number of combinations is obviously 45 over 10, that's the overall number of ways you can arrange children in any order. So you can pick any 10 children, and then you simply do the division. So the probability of getting 3 girls is 15 over 3 multiplied by 30 over 7, that's the corresponding probabilities for the boys and the girls, and they are divided by the overall number probabilities. And once again, you get a value, in this case, 0. 29.

Multinomial Coefficients

Having discussed binomial coefficients, we're now going to discuss multinomial coefficients. So, this is the idea of having not just two groups that you want to separate your elements into, but actually having several groups, more than two. So for example, let's suppose you have 10 students in a class and you want them to form 3 different groups consisting of 4, 3, and 3 members, respectively. Now the question is, how many ways can you arrange the students to belong to

these particular groups? So, you have to take slow when you are considering these kinds of problems. So first of all, if you take the first group, you actually want to choose 4 students out of the 10 students which are available, and that's obviously 10 over 4 arrangement. And then you are left with 6 students, and the number of ways to split 6 students into groups of 3 and 3 is 6 over 3, respectively. So that kind of finishes off your calculation, so it's a multiplication in this case of the choices made when you're picking the first group and the choices you made picking the second group. The third group is kind of arbitrary here, and what we end up with is we end up 10 over 4 multiplied by 6 over 3, which after you perform the factorial calculations, you get a value of 4, 200. Now what's important here is not so much the final value as the numbers I've highlighted in blue. So as you perform the calculations you're going to see that some of the terms in the denominator of the first element are going to be in the nominator of the second element, numerator rather, and this is going to continue. So if you have more than three groups, for example, you're going to have these additional cancellations of terms throughout this entire thing, and this certainly simplifies the process of calculating multinomial coefficients. So, speaking of multinomial coefficients themselves, in general what we have is we have the following situation. So we have, when you have a scenario where you are splitting elements into several different groups of a particular size, you have a product of these n over n1, n - n1 over n2, and so on, and here what we're doing is we're saying there is k number of groups, so there's k different groups, and they all have size n1, n2, all the way up to nk. And so the calculation we'd use is, using those blue terms that I showed in the previous slide, it reduces to n factorial divided by the product, so the factorials of the group sizes. Now this is also something called the multinomial coefficient, and it's also written like this. So on top you have n and on the bottom you have the sizes of the actual groups, and just like with the binomial coefficients it actually shows up as a series expansion for a sum of terms taken to the power of n. So here are a few examples of calculations involving multinomial coefficients. So first of all, if you go back to the example where we had 10 students and we picked groups of 4, 3, and 3, we get 10 over 4, 3, 3, that's 10 factorial divided by 4 factorial, 3 factorial, 3 factorial, which is, once again, equal to 4, 200, that's the same exact result that we got in the previous calculation. Now, here's another one. Let's suppose that you have 3 As, 4 Bs, and 5 Cs and you want to calculate the number of ways in which you can actually arrange these. Now, that would be 12 over 3, 4, 5, so once again you take the factorial of everything, you make a fraction, you divide and you end up with 27, 720. And of course, if you're into algebra you can also calculate expansions and coefficients of terms in that expansion. So for example if you have x plus y plus z to the power 3, then the coefficient of the term x squared z, which you would have to rewrite as x squared, y to the 0, z to the 1, so that's 2, 0, and 1 in the powers, that would be a multinomial coefficient, 3 over 2, 0, 1, which is equal to 3.

Probability of a Union of Events

The last thing that I want to mention is this idea of a probability of a union of events and how this calculation can actually help us figure out certain problems. So, from our definition of probability, we know that for a set of disjoint events, so if you have a sequence of events A, like A1, A2, A3, and so on, and they don't overlap, then the probability of a union of these events is the same as the sum of probabilities. Now, for every two events, and not necessarily disjoint events, we can also formulate the following, we can state that the probability of a union is equal to the sum of probabilities minus their intersection. And the way you can think about it is as follows. If you think of a union as the circle A added to the circle B, then if you want to actually get the area of the shaded-in area, then taking the area of A and area of B and adding them up together yields you double the intersection, so that's why we have the first term plus the second term, but then subtracting the intersection. And this same approach can actually be extended to three or more events. So we can have the probability of a union of three events, and in this case you would add up their respective probabilities, you would subtract the sum of their intersections, but then of course you've subtracted that central overlapping piece one time too many, so you have to add it back. Now these formulae might look a bit strange, so let's take a look at how you would actually use them. So here is a small scenario. If you consider a cohort of, let's say, 200 students, let's say that 50 students take programming, 100 students take electronics, 150 students take math, so these you can consider as kind of ordinary sets which don't overlap anything, then we can say that 30 students take programming and electronics, 45 students take electronics and math, and 25 students take electronics and programming. Furthermore, we can state that 15 students take all 3 classes, and then we know that some students take no classes at all, so some students aren't even part of this whole mathematics, electronics, programming kind of routine, they're doing something else. Maybe they're doing languages or biology or something else. And what we want to find out is we want to find the probability that a student, if you take 1 of these 200 students, we want to find out the probability that the student takes at least 1 class, so the student is actually part of those students which are doing all of these science-y things. Of course, we're going to use this idea of probability of a union of events to calculate exactly that. So first of all we calculate the simple probabilities for programming, engineering and math, respectively. So if you pick out a student, the probability, for example, that they take engineering is one-half, because it's 100 over 200, so we get these values. Then we get the intersection values, so the probability of a student taken at random taking both programming, as well as engineering is 3 over 20, so that's 15% effectively. And finally, we have that intersection of the students overall, so the students which take all subjects, they're 15 out of 200. And now we have that formula that we looked at in

the previous example where you can calculate the union just by having all of these values here. So we can effectively calculate the union, which would imply that a student actually takes at least one of these, or maybe more, because that's what we're trying to find out here. So that's the sum of the individual probabilities minus the sum of the intersections, but because there's 3 elements we have to add up the overall intersection, and that gives us a value of 140 over 200, or just 0. 7. So the probability of a student taking either programming or engineering or mathematics is in fact equal to 0. 7. Of course a student can take all three of them and they will still be part of this group.

Summary

Alright, so let's try to summarize what we've learned as part of this module. So, to study different phenomena we first of all perform experiments and we record our observations. Now, the sample space describes all the possible outcomes that we expect from an experiment, and this sample space can be modeled using set notation and it can be measured using the different counting methods or combinatorics that we've looked at. In addition, we talked about this concept of probability, which is a value somewhere in the range from 0 to 1 inclusive, and it quite simply describes the likelihood of an event actually occurring. And finally, we looked at this idea that if events are mutually independent, the probability of a union of these events is the sum of their individual probabilities, and that is a central concept upon which many of probability calculations are actually built.

Calculating the Conditional Probability of Events

Overview

Hi there, and welcome to this module on calculating the conditional probability of events. So the goal of this module, there's only one goal, is for you to understand the concept of conditional probability and learn how to apply it to particular problems. So here's what we're going to see in this particular module. First of all, we'll obviously have to discuss what conditional probability actually is, then we'll discuss the idea of dependence or independence of individual events, we'll talk about the different laws of conditional probability, we'll also discuss the idea of partitions,

and we'll use the idea of partitions in our discussion of the Law of Total Probability. Then we'll take a look at the Bayes' Theorem, and finally, we'll take a look at one particular problem, which is very relevant to conditional probability, the Gambler's Ruin Problem.

Conditional Probability

So let's first of all discuss the ideas of conditional and unconditional probability. Let's imagine that, for example, you're interested in outdoor sports, maybe you play outdoor sports yourself or maybe you like to bet on sports. Now one of the things you might care about is the probability of rain, and there are different ways in which you can actually talk about this probability. So one way to figure out the probability is to just try to guess the probability it will rain today given no additional information. So let's suppose you're sealed in a bunker somewhere, you have no idea about anything going outside in the world, and you want to find out the probability it will actually rain today. Now compare this with the probability that it will rain today given that it's been raining an entire week. Now, I'm sure you will agree that in this particular case the probability should be much higher because you already have some statistical information about what's been going on through the entire week, and so it's much more likely that it will rain today, and you can call it a rainy week or a rainy month, or whatever. So the probabilities are different, and when we talk about these probabilities we say that the first value, the value where you have no additional information, is the unconditional probability, and similarly for the second value, we call it the conditional probability, because it's actually conditional upon other observations that have already been made. So here is the formal definition of conditional probability. So the conditional probability of an event A, given an event B, is recorded as capital P, and then round brackets, and inside we have A vertical bar B. So we read it as P of A given B, or the probability of event A, given event B, and it's defined as the probability of the intersection of A and B divided by the probability of B. And of course we are making the assumption here that the probability of B is greater than 0. So let me show you a practical example of conditional probability using the formula that we've just seen. So for example, the unconditional probability that you take a 6-sided die, a balanced die, and your roll a 3 on it is one-sixth, the same as if you were to roll a 1 or a 5, for example. Each side, provided it's a balanced die, has a probability of one-sixth. But what if I tell you in advance that I rolled a die and I observed an odd number, and I'm asking you to figure out the probability that it was a 3. Well, this reduces the number off different options, because I just told you it's an odd number, so it can be 1, 3 or 5, and as a result, this probability, the probability that I've rolled a 3, turns from an unconditional probability to a conditional probability. It's contingent upon the fact that I rolled an odd number, in this case, and so the conditional

probability here is one-third. Hopefully that is obvious because we only have three different options. If I tell you it's an odd value, it can be 1, 3 or 5, so the probability it's a 3 is exactly one-third, but we can actually figure this out, not just by intuition, by counting the number of samples, but by using that formula that we've seen in the formal definition. So using our definition of conditional probability, we can calculate the probability that it's a number 3 given that it's odd, and this is the probability of 3 intersected with odd, and of course the probability of 3 intersected with odd is the probability of 3, which is one-sixth. And similarly, the probability of an odd number is one-half because half of them are odd and half of them are even. So using that formula, we have the probability of rolling a 3 given that the number is odd is one-sixth over one-half, which is one-third, which is exactly what we get through our intuition, so to speak.

Independence of Events

Now we need to talk about something called the independence of events. We typically say that two events, A and B, are independent if any of the following three conditions hold. So either the probability of A given B is the same as the probability of the A or the probability of B given A is the same as the probability of B, or the probability of their intersection is the product of their probabilities. So if none of these conditions hold, then the events are dependent, but if at least one of these conditions holds, then we can say that the two events are in fact independent. So let me show you how you can use this definition to calculate the dependence or independence of particular events. And once again we're going to talk about die rolls, so we're going to have events generated by a single, balanced, six-sided die. So let's suppose we have event A, which tells us that an odd number was rolled, we have event B, that an even number was rolled, and we have event C, which is defining that somebody rolled either a 1 or a 2. Now the question is, are events A and B independent? Well, we know that the probability of A is one-half, and the probability of B is one-half. Now their intersection, the intersection of these 2 events, is actually 0 because you cannot have both an even number and an odd number at the same time, and the conditional probability of A given B, and similarly the conditional probability of being given A is actually equal to 0. And because the probability, the conditional probability, is not equal to the original probability, and the intersection is actually an empty set, we can say that these events are in fact dependent. Now another question is are A and C dependent, or indeed independent? Now the probability of A given C is equal to one-half, and that also happens to be the probability of A, and as a result we can say that these two events are in fact independent. Let's take a look at a more complicated example with regards to calculation of independence of events. So we're going to set up a much more complicated scenario. So this time around, let's suppose you have three

brands of wine called X, Y and Z, and they're being ranked by judges. So we have judges, they taste the wine, and they set up rankings for that wine, and we have different events. So we have event A, which specifies that the wine X is ranked higher than Y, we have event B, which specifies that wine X is ranked the best, event C that wine X is ranked second best, and event D that wine X is ranked worst. Now once again what we can do is we can represent the entire sample space because even though this problem is challenging, the sample space isn't that big. So here I've listed the six possible results in terms of rankings. So for example, E 1 represents the fact that wine X was ranked highest, and then Y and then Z. Now we can start talking about the possible outcomes for our observations. So for example the event where wine X is ranked higher than Y, we can see that it's E 1 E 2 and E 5 in this case, we can for event B see that X is ranked best in E 1 and E 2, those are the elements which begin with an X, then we have event C, which has E 3 and E 5, and finally we have event D, which has E 4 and E 6 in it. So now what we can do if we want to find out the independence of these events is we can calculate the probabilities. So first of all the probability of A, the probability of the wine X being ranked higher than Y is 3 over six, you can see that the set A contains 3 elements out of the possible 6, so that's 3 over 6, that's one-half. The probability, or the conditional probability now of event A given event B once again applying that formula, you'll see that the probability of A intersected with B, if you look at sets A and B, their intersection is in fact the set B. So you have probability of B divided by probability of B, which is equal to 1, we have the conditional probability of A given C, that's a more interesting setup, the probability off A intersected with C divided by the probability of C, and of course if you intersect A and C, the only element you get is E 5, which has probability of one-sixth, because remember we have a total number of 6 events, so that's one-sixth of one-third, which is one-half, and finally, the probability of A intersected with D, or the probability of A given D is defined as the probability of A intersected with D divided by the probability of D. But if you look at A and D, they don't actually overlap, they don't have any elements in common, and so this probability is actually equal to 0. And once again from the definition that we already have, what we can figure out is that events A and C are independent because you can see that the values here are identical, but we can see that events A and B are dependent, and similarly events A and D are dependent as well.

Multiplicative and Additive Laws

So now what we're going to take a look at is some Laws of Probability. We're going to take a look at the Multiplicative Law and the Additive Law of Probability. So let's start with the Multiplicative Law. We're going to begin by looking at the formula for conditional probability that we already

looked at, and essentially we have this formula that the probability of A given B is equal to the probability of the intersection of A and B divided by the probability of B. So what you can do is you can rearrange this formula, and you have the probability of A intersected with B being equal to the probability of B multiplied by the probability of A given B, and because the intersection terms can be swapped the other way around, this is also equal to the probability of A multiplied by the probability of B given A. Now if these events are independent, then all you really have is a simple product. So if events A and B are independent, the probability of the intersection of those events is simply the product of their individual probabilities. So let's take a look at an example where we can use the Multiplicative Laws. So let's imagine a scenario where we're rolling a six-sided die and we can generate two different dependent events. So we can roll an odd number or we can roll a number which is less than 4. Now we can calculate the probabilities of the different events associated with this. So the probability of A, the probability of getting an odd number is one-half, because half of the numbers on the die are even and half are odd, so the probability of odd is one-half. Now the probability of having a number less than 4 is also equal to one-half, because half of the numbers are 4, 5, and 6, the others are 1, 2, and 3, and that's what you get in B. Now we can also calculate the conditional probability. So the conditional probability of A given B is equal to two-thirds, essentially this is the conditional probability that you have an odd number given that you know you have a number of less than 4. If you have a number less than 4 that's 1, 2, and 3, and out of 1, 2, and 3, 2 out of 3 are odd, so that's why you get two-thirds, and the other probability is is also two-thirds, because if you do it the other way around, you also get two-thirds. And so now we can apply this law, so the probability of an intersection is the probability of A multiplied by the probability of B given A, which is one-half multiplied by two-thirds, which is equal to one-third. Now if you think about it, if you think about just the intersection of A and B, so A is the odd numbers, so that's 1, 3, and 5; B is a number less than 4, which is 1 and 3, so you have the numbers 1 and 3 out of a total of 6 numbers, so that's exactly one-third, and as you can see this result agrees with our intuition in this case. Now let's talk about the Additive Law of Total Probability. So in the previous case we had the intersection, and now we're going to have the union. So the probability of a union of events, and this is something that we've met before, is the sum of the individual probabilities minus their intersection. Now if A and B are mutually exclusive events, meaning their intersection is 0, then the probability of a union is quite simply the sum of the individual probabilities. Now both the Multiplicative and Additive Laws can also be extended to calculate the probabilities of more than two events. It does get a bit more complicated, but you're welcome to try it out on paper on your own time. I want to present those results here.

Law of Total Probability

Now we're going to talk about something called partitions. Now it's a very simple idea of what you see here on the left is a sample space, and I've split it up into five different parts. So if S is the sample space of some experiment, given K disjoint events, so events which don't overlap, and we can call them B 1, B 2, B 3, and so on, the union of these events actually form the sample space. And what we say in terms of the terminology that we use, we say that these events form the partition of the sample set S. Now the reason why we're discussing partitions is to talk about the Law of Total Probability. So let's suppose that we have these partitions, let's call them B 1 all the way up to BK, they're partitions of the sample space S, and the probability is greater than 0 for every single one of these partitions, then for every event A in S, what we can say is the probability of A is the sum of the conditional probabilities of A given that partition, given that we are in that partition multiplied by the probability of that partition. So we're essentially summing up the product of the conditional probability of A being in that partition multiplied by the probability of the partition itself. If you think about it, it's very logical, it makes a lot of sense. There is also a conditional version if you have the event A conditional on some event C, and it's a similar scenario, except that in the case of the first product we have B J conditional on C, and in the second case, in the second product, we have a conditional on the intersection of B, J, and C. So this is the conditional version of the Law of Total Probability. Now the Law of Total Probability, once again, may seem a bit cryptic and a bit bewildering in terms of how to apply it, so let's take a look at an actual example where we can try to apply this law. So here is a scenario. Let's suppose that you have a player, and the player plays some kind of game where they get a final score once they've played. So the final score is somewhere from 1 to 50. Now the player's score after the first game is equal to X, and the player continues to play the game until they obtain another score, such that Y is greater than or equal to X. So they obtain the score Y and they continue to play while Y is greater than or equal to X. So essentially they continue to play so long as they are winning, and the question is what is the probability that they reach 50? What is the probability that they actually reach the highest score that is available? So let's make a definition, let's define B I to be the probability that X has the value I. Now conditional on this B I, the value is likely to be any number, so it can be I, I plus 1, I plus 2, and all the way up to 50, and each of these 51 - I values for Y are equally likely. So what we can say is we can say that the conditional probability of event A given B I is equal to the conditional probability that Y is equal to 50, because that's our definition of A given that B I, and that is equal to 1 over 51 - I. Now what we can say is because the probability of B I is one-fiftieth for all values of I, we can calculate the probability of A, which is, once again, this is where the Law of Total Probability comes in, so it's

the sum of this product. So essentially we have this sum going from 1 to 50 of one-fiftieth divided by 1 over 51 - I, and you can take out the one-fiftieth, and what you end up with in round brackets is you end up 1 plus one-half plus one-third, and so on, and all the way up to one-fiftieth, which is equal to 0. 09. So that's our final value, that's the probability of actually reaching a score of 50.

Bayes' Theorem

All right, so we're now going to take a look at Bayes' Theorem, and we're going to start by looking at an example of where Bayes' Theorem can be applied. So let's imagine that you have a medical test where you can test the person for the presence or absence of a particular disease. But this test isn't perfect, it's not foolproof, instead it's kind of 90% reliable, and what we mean by that is that if a person has the disease, the probability that they will test positive is 0. 9, and if a person doesn't have the disease, the probability they will still test positive is 0. 1, which isn't particularly encouraging. Now we also know from the environment, from the statistics that are out there, that the chances of actually having this disease are fairly low; they're 1 in 10, 000. Now let's suppose you go to a clinic and you do this test, and unfortunately you test positive. So the question is, well, what is the probability that you actually have the disease? We're going to be using Bayes' Theorem for this one, so here it is. So essentially in the Bayes' Theorem, what we do is we define events B 1, all the way up to B K, which form a partition over the sample space such that the probability of each of these events is greater than 0, and we also define event A, such that, once again, the probability is greater than 0, and then for every single one of these partitions, we have the following relationship. So the probability of B I conditional on A is this particular formula, the product of B I multiplied by the product of A conditional on B I divided by the sum, going from 1 to K of the probability of B J multiplied by the probability of A given B J. And there is also, once again, a conditional version of Bayes' Theorem that you can see down below, all you do is you simply kind of intersect the events. So if you want the probability, the conditional probability in fact of B I given A intersected with C, that's the output that you get on the right. So let us now apply Bayes' Theorem to our problem of testing for diseases. So what do we know? Well, we know that the probability that we have the disease is definitely not 0. 9. You might be tempted to think that, but we cannot really discount the information that only 1 in 10, 000 people actually has the disease in the real population. If we were to discount this information, we would have an unconditional probability value, but we do have this information, so we're bound to calculate conditional probability instead. Now let's have B 1 and B 2 denote the partitions, if you will, that we have the disease, and we do not have the disease, those 2 form a partition over the overall sample space, and we're also going to have A denote the event, the

response to the test was positive. And so what we're really trying to do is we're trying to figure out the probability of B 1 given A. With the Bayes' Theorem, what we can do is we can take all those probabilities and we can plug them right into the formula that we just looked at, and we get a concrete value. We get a value of 0. 0009, so that's almost 1 in a 1, 000, so the conditional probability that we actually have the disease is not 0. 09, it is in fact only about 1 in a 1, 000. Now, here's how you can interpret the result. So only 1 person in 10, 000 actually has the disease, but the test gives a positive response in 1 person in every 10, so there is a mismatch. As a result, the number of positive responses is 1, 000 times more than the number that the people have the disease. So, the whole system has an error margin of 1, 000 fold. One out of every 1, 000 people for whom the test gives a positive response actually has the disease, and that's the result that we're getting here.

Gambler's Ruin: Problem

All right, so now we're going to talk about a particular problem in statistics, and that problem is the Gambler's Ruin Problem. It is actually a very popular problem for discussing the outcomes of a repeated series of bets, and the earliest mention of this problem was a letter from Pascal to Fermat in 1656, so quite an old problem. Now here's what the problem is all about. Essentially you have a setup where you have two gamblers, A and B, and they're playing some kind of game against one another. The actual game isn't important, what's important is that on each play of the game, there is a certain probability that gambler A will win $1 from gambler B, and there is that 1 - P probability, the reverse probability that gambler B is the one that's going to win $1 from gambler A. Now initially we have gambler A has a certain starting hand, they have a certain balance or a certain fortune of I dollars, and gambler B has a fortune such that if you add the two fortunes together, you get some value K. So you can think of it like a game of poker where there's a certain fixed number of chips on the table, nobody adds new chips, and those chips are simply being redistributed until eventually people ran out of those chips, and whoever runs out obviously loses, so it's the same idea here. You essentially have gamblers which keep playing the game repeatedly and independently until one of them reaches a balance of 0, and that gambler is effectively ruined. So what can we say about this game? Well, we can consider a game from the perspective of player A, that's how we're going to simulate it in actual fact. So player A has an initial balance of I dollars, and depending on the probability, the game might be favorable or unfavorable to that player. So for example, if the probability is greater than one-half, then obviously the game is favorable, and if the probability is less than one-half, the game is unfavorable. If the probability is exactly one-half, then the game can be considered fair or

balanced. Now the game ends when one of two things happen. Hopefully you've already realized what those two things are. So either the player A reaches K dollars, so player A gets the whole pot, all the chips or all the money that was initially in the play, and this means that player B has a balance of 0, or the reverse happens, so player A gets ruined, they reach $0, and player B has all the money. Now the problem here, and it's not an easy problem to solve, is to determine the probability that the fortune of gambler A will actually reach K before it reaches 0. So it's not a simple problem of calculating just one step of the game and telling us what's going to happen on the next run of the game, but actually trying to figure out the entire sequence, because well, this game is called Gambler's Ruin, it's all about this idea that one of the gamblers must eventually go bankrupt, and so you want to have a probability value for how likely you are not to go bankrupt, how likely you are to actually win the whole pot, and that's what we're going to take a look at. The solution isn't particularly simple, but we're going to take a look at it now.

Gambler's Ruin: Solution

All right, so now we're going to take a look at the solution to the Gambler's Ruin Problem. So we're going to introduce lots of notation here to represent the different states of the system. So first of all we have small A with a certain subscript I, and this is going to represent the probability that gambler A reaches the balance K, so they get the whole pot before getting ruined given the initial fortune of I. So here the subscript actually indicates the starting fortune of gambler A. Now in addition, we're going to have for each value J, going from 0 to K, so each time we observe a set of plays that lead to A's fortune reaching the value J, The conditional probability that A wins is actually going to be a subscript J. Now if gambler A's fortune reaches 0, then A is obviously ruined, so A of 0, A subscript 0, is actually equal to 0. Now we'll introduce even more notations that will have capital A and capital B. So capital A with the subscript 1 will denote the event that A wins $1 on the first play, and similarly B 1 will indicate the same for gambler B. Now we'll have W denote the event that the fortune of Gambler A reaches K before reaching 0, so that's what we are actually after, we're after the probability of W, and we can represent it as follows. So the probability of W is the sum of the probability of A 1 multiplied by the conditional probability of W given A 1, plus the probability of B 1 times the probability of W, given B 1. Now of course we know that probability of A 1 is actually P, that's our P value, on the probability of B 1 is 1 - B, so we can substitute to those values, and we have the expression that you see here. All right, so the initial fortune of gambler A is I dollars, where I can go from 1 to K - one because of the other 2 states, the 0 and the K, they're kind of the terminal states. And the probability of W is represented as a subscript I. Now if a gambler, A, wins $1 dollar on his first play, then his fortune becomes I plus 1,

and the conditional probability of the win given A 1 can be represented as a subscript I plus 1. Now if A loses $1 on the first play, then his fortune becomes I - 1, so he lost $1, and the probability of W given B 1, because remember, A 1 lost, so B 1 is kind of a loss for player A is a subscript I - 1. So we have the following representation. We have A subscript I as probability that the P value multiplied by A subscript I plus 1 plus 1 - P multiplied by A subscript I - 1. And we know the terminal state. So we know that A 0 is equal to 0, we know that A K is equal to 1, so if we take all the other values, we have the following set of relationships. So we have A 1 expressed in terms of A 2, we have A 2 expressed in terms of A 2 and A 1, and all the way up to A subscript K - 1, which is represented as simply P, because remember, A K actually equal to 1, so A K is kind of disappearing here, plus 1 - P of A K -2. So we have this kind of recurrent relationship if you will between the different terms. Now what we can do is we can take each of the terms here and we can rewrite a subscript I as the sum of P times E subscript I plus 1 - P times A subscript I, and then we can rearrange things and we get the following. So, as you can see what's happening here is you're starting out with A 1 being equal to PA 2, and then of course what you can do is you can use this transformation, you can rewrite the left-hand side, so A 1 can rewritten as PA 1 plus 1 - PA 1. And we have the same thing on the right-hand side. Then we can move PA 1 over to the right, so we end up with 1 - P times A 1 equal to PA 2 - PA 1, and then rearranging, dividing everything by essentially P, the right-hand side can be moved left, the left-hand side can be moved right, and so we have A 2 - A 1 being equal to 1 - P over P multiplied by A 1. And we can do the same for every single one of the other terms, and here's what we end up. And this is actually going to help us simplify things quite a bit. We end up with the following set of relationships. So you can see that in each of the different cases, what we have is we have A subscript something minus A subscript, that thing -1, being equal to 1 - P over P taken to some power multiplied by A 1, and this relationship continues in each one of the terms. And hopefully you can already see the trick here. What if you add all of these up. Well, if you take the sum of the left and right-hand sides, you'll notice that on the left you're going to have massive cancellation. You're going to have A 2 cancel with - A 2, A 3 cancel with - A 3, AK - 1 cancel with - AK - 1, and so what you end up with on the left is very simple, you end up with 1 - A 1 being equal to A 1 times the sum of terms, and here we have 1 - B divided by P to the power of I, with I going from 1 to K - 1. And this is our solution, this is the solution to the Gambler's Ruin problem. Now having acquired the solution, we can actually plug in this formula for both fair and unfair game. So let's take a look at the fair variety first of all. So, P is equal to one-half, and remember, this is our formula, this is the formula that we came up with. So what we end up with, because if you plug P equal to one-half, you get 1 - one-half divided by one-half, so that's obviously 1 to the power of I, and it doesn't matter what I is because 1 to any power is I, and so you end up with a very large simplification because you end up with 1 - A 1

being equal to A 1 times K - I, and so A 1 is equal to 1 over K, and you can simply start plugging in these values. So for example suppose that A has a balance of $98, B has a balance of $2. In a fair game, so in a game where P is equal to one-half, to calculate A subscript 98, you simply divide 98 by K, and of course K is the sum of the fortune, so 98 plus 2, that's 100, so you have A subscript 98 being equal to 98 over 100, which is 0. 98. Okay, so now let's take a look at an unfair game, so let's imagine that we have a probability, which is not equal to one-half. Then what we can do is we can once again change the sum a little bit. So here is our result, the result that we acquired, and we can rewrite it in a slightly different form. We can rewrite it in the following form. So essentially this is a slight rearrangement of what we have above, but we end up with an expression for A 1, and this expression for A 1 is where we can subsequently replicate this expression for the different subscripts for I equal 2, and all the way up to K - 1, and as a result what we end up with is we end up with the following. So essentially we have this expression for A subscript I in terms of 1 - B over P to some power, and we can plug in the values in here as well. So for example let's suppose that P is equal to 0. 4, A has a balance of $99, B has a balance of $1. So essentially, first of all, we can calculate 1 - P over P, that's 3 over 2, I is equal to 99, K is equal to 100, so if we plug in those values, you'll see we have 3 over 2 to the power 99 - 1 divided by 3 over 2 to the power 100 - 1. Now we can sort of neglect the - 1 because it's not particularly significant compared to the other terms, so if we ignore the - 1, we just cross them out, we have 3 over 2 to the power of 99 divided by 3 over 2 to the power of 100, which is approximately two-thirds, so this is our final value. And notice that this value is significantly different from simply the value of 1 - B, because even though you might think that there is only 1 step from getting from $99 $100, and in the game there is also a probability, and it's a significant probability, that at $99 player A is first of all going to lose some money and go down to maybe $98 or 97, or whatever, and then come back all the way up to $100, and this is why the value is two-thirds and not some other value like 0. 4 for example. You might say that, well, there's only one step, so the probability is that well no, in actual fact we don't know how many steps there are to reaching $100, and the probability value certainly reflects that.

Gambler's Ruin: Simulation

All right, so what we're going to do now is we're going to try and see whether we get the same values from simulating Gambler's Ruin as we do from our theoretical result. So I'm going to write the simulation for Gambler's Ruin, and it's actually a very simple simulation. Now first of all I have to point out that we cannot just simulate a sequence of games at once, we need to do it more than once, because on a single run, the value that we're going to get is not going to be precise enough. So what I'm going to do is I'm going to specify the number of simulations to be, let's say

1, 000, and also I'm going to make a logical array of values for the instances of player 1 winning or losing, so I'm going to have wins being a logical array, and the length of that array is going to be the number of simulations. So each of the elements in this survey is going to be true if player 1 won, and false if player 1 lost. So now we need to specify the parameters of the actual game, the probability value, as well as the balance of player 1 and player 2 when the games actually begin. So the probability value will be something like 0. 5 to begin with. The player 1 balance is going to be, for example, 98, the player 2 balance is going to be 2. There we go. Okay. So, what we do now is we simulate the game sequence for however many times we chose, which in our case is 1000. So I'm going to have a for loop, so for i in 0, all the way to the number of simulations, and should be an S here at the end, just add an S here. So for every single one of these simulations, here's what we do. We initialize local variables for the balance, so p1b becomes p1_balance, and the same for p2b, p2_balance, and then what we do is we make a while loop which keeps going until somebody gets ruined, until somebody goes bankrupt. So I'm going to have a while loop. So while p1b is greater than 0, and p2b is also greater than 0, then we perform the actual simulation. So the simulation itself is very simple, I just generate a random number and I see if it's less than the P value. If it's less than the P value, then player 1 wins; otherwise, they lose. So we say p1_wins is equal to random uniform number, so runif, just a single number, and it has to be less than the P value. So that's how we determine whether player 1 actually wins or loses, and if player 1 wins, then we adjust the balances. So we say p1b equals p1b + 1, and p2b = p2b - 1, otherwise we do the exact opposite. So we say p1b = p1b - 1, and p2b = p2b + 1. There we go. So this is how we simulate the game, and of course at the end of the game when somebody gets ruined, we need to record it, and we have our wonderful logical array called wins for that purpose. So we say wins at i is equal to, and we simply compare p1b with a value of 0. So if p1b is greater than 0, that means player 1 won, and we put a true in there, otherwise we put a false in there. So having made the simulation, what we can now do is we can sort of paste in the results of that simulation and see the probability of the player actually winning. So here I'm going to use the paste command and I'm going to say, player 1's probability of winning is, and then I'm going to basically sum up all the wins, and it's a logical array, so as we sum it up it becomes essentially false is a 0, true is a 1, so we get a sum, and we divide it by the number of simulations. Let me just put it in round brackets here for extra conciseness. There we go. So this is our Gambler's Ruin simulation. Now let's remind ourselves of the theoretical results that we got as we looked at the solution to Gambler's Ruin. So we had two different examples. We had a fair game, and in the fair game we had exactly the conditions that I've entered right now. So the probability was 0. 5, the balance was 98 and 2 respectively, and we're expecting the value somewhere around 0. 98. Now the second game was an unfair game where the probability was 0. 4, the balances were 99 and 1, and

we're expecting two-thirds. So we're expecting about 0. 66, however close we get to that is another question. Obviously we are only doing a 1, 000 simulations, so we won't get precisely that number. Okay, so we're going to start by running the fair game simulation. So we're expecting 0. 98, let's see what we actually get as we run all of this. So let me bring up our interactive here, and this you can see the value here is 0. 981, and we can sort of keep trying that again and again and getting different values, so now it's 0. 986, and we can go again, and now it's 0. 983. So it was reasonably close to the expected value of 0. 98. Okay. So coming back to the source code, now what we can do is we can plug in the values for the unfair game, so that would be 0. 4, 99, and 1, and once again I'll save this and I'll open our interactive, and we can now source this entire thing with Echo and see the values we get this time around. And notice this 0. 669, and we're expecting about two-thirds, we were expecting 0. 66 and 6 to infinity, so we're certainly getting something which is very close to what we had actually expected. Now we're getting 657 here, and now we are getting 0. 686, so we are getting the right numbers, and this helps validate our theoretical results to some degree.

Summary

All right, so let's try to summarize the things that we've learned in this module. So this module was all about conditional probability, and conditional probability of an event is simply the probability of an event given some prior information, which can actually affect this probability. We also saw this idea of independence. We saw that events can generally be tested for independence, and the rule is very simple, independent events do not influence each other's probabilities, and you can look at conditional probabilities and see if the values match the ordinary probabilities, and if they do, that means the events are independent. We also discussed something called Bayes' Theorem. This is a theorem that describes the probability of an event based on prior knowledge of conditions related to that event. And finally, we looked at the Gambler's Ruin, which is a problem, which is a good model for predicting the eventual outcomes of a series of repeated bets.

Understanding Random Variables and Distributions

Overview

Hi there. In this section of the course, we're going to take a look at random variables and their distributions. So the goal of this section of the course is fairly simple. I want you to understand this notion of a random variable, and we're going to take a look at some of the common distributions of random variables, how they behave, and what they're actually used for, or why would you want to even discuss them in the first place. So here is what we're going to see in this section of the course. First of all, we'll discuss the obvious part, which is what is a random variable and why do we care about them? We'll also discuss the difference between discrete and continuous random variables. Then we'll discuss some of the concepts associated with random variables, such as the idea of distributions and probability functions. And then we'll discuss the two classes of random variables. So first of all, we'll discuss the discrete distributions. So uniform, binomial, geometric, and hypergeometric, and then we'll discuss the continuous distributions, the uniform distribution, the very popular normal distribution, as well as the gamma and beta distributions.

Random Variables

All right, so we have to begin by discussing the definition of a random variable. There are actually two definitions. There is the formal and informal definition. So the formal definition, which isn't going to help us much, is that a random variable is a real value function on the sample space. So what does that mean? That means that basically for every single element of the sample space, you can feed it into some function and get some value out of it. So that's one way of looking at things. But the informal definition is probably a lot simpler because it's just human-readable. So, a random variable is quite simply a variable that can take on a random value, and this value can either be taken from a finite set of values or an infinite set of values. So let's talk about the difference between discrete and continuous random variables. So a discrete random variable can take on a set of values which is either finite, so for example you have something like a coin toss that's only two possibilities, heads or tails, or you can have a die roll which gives you six possibilities, or the set of values can be countably infinite. So that's probably a more complicated idea to kind of understand. The idea of countable infinity is basically when you have something that can be put into a one-to-one correspondence with natural numbers, so natural numbers, remember, are whole, positive numbers, if you can have something that has a one-to-one correspondence with natural numbers, we call it countably infinite, and that is also treated as a discrete value. So one way of talking about the countably infinite value is as follows. Let's suppose that we decide to measure people's height, for example. We round a person's height to some value, like let's say if you measure it in centimeters, you round next to the next centimeter.

So you're basically saying that if somebody's 200. 578, whatever, centimeters tall, you just round it to 200 basically. So in this case, we still have plenty of possibilities, because on the one hand you could argue that nobody's taller than 3 meters, so you have 300 distinct cases. But on the other hand, you may end up something even taller. Maybe you're measuring buildings, for example, in which case there is no upper bound. So in this case we say that yeah, it's a discrete value, but there is a countably infinite amount of values that actually exist. So that's one side. That's the discrete random variables. And then there is the continuous random variables, which can take on an infinite set of values. And speaking of people's height, we can talk about a person's exact height right down to the molecule. Now interestingly enough, if you try to measure a person's exact height, well, first of all, it's impossible virtually, how do you measure next to the sort of molecular size? But this yields an infinite number of values. So it's not a finite set of cases, it's virtually infinite. So the probability that you or any other person watching this video, or out there in the world, is exactly 2 meters tall, as in 2 meters and 0 molecules, or atoms, or whatever on top of that is equal to 0, because you'll never find anybody who is exactly 2 meters down to the smallest iota, so to speak. So there are ways of treating this value. One way is we talk about continuous random variables. We're going to have continuous distributions later on. That's the mathematical apparatus for dealing with such variables. Or you can turn it into a discrete value just by rounding. So for example, you just take everybody's height, round it to the nearest centimeter, for example, and that's it, you now have a discrete set of values. You can plot histograms or whatever with them. So that's an alternative approach. All right, now let's talk about the notation for random variables. So random variables are typically denoted using a capital letter such as capital X or a capital Y. And when we want to record this idea of a probability of this random variable taking on a specific value, here is how you write it. So for example, let's suppose that we are talking about die rolls and we want to find the probability that this random variable X, which represents the result of a die roll, is equal to 3, so that's how we write it. We write the letter P followed by round brackets, and then we say capital X = 3, and we give it a value. So the probability of X having the value of three is one-sixth. Now generally, the probability of a random variable, X, taking on some value, small x, is expressed as follows. So you have capital P, once again, and then you have capital X = small x. So this might seem a bit confusing, but just remember, the capital letter represents the random variable, and the x here represents a possible value that this random variable can actually take. So in our case, I'm saying that the probability that X has the value small x is equal to 1 over x squared, for example. And this can, in fact, be a function of x, as we'll see later on.

Discrete Random Variables

All right, so let's talk about discrete random variables for a second. So essentially, a discrete random variable is a variable that takes on a finite or countably infinite set of values. Some of the examples, well, we looked at some of them already. So for example, a single coin toss will give you either heads or tails, or you can toss the coin 10 times and count the number of heads in 10 different coin tosses. That is also a finite set of values that can be generated from that. Once again, a die roll gives you 6 possible values, and you can roll the die 100 times and calculate the sum of those rolls. That would also be a discrete random variable. Or something different, something like a person's ranking in a competition, because if you have n people in a competition, for example, their ranking would be first, second, third, all the way down to nth. So once again, it's a finite set of cases. Now I have to stress that in all of these situations when we talk about discrete random variables, they don't all have to be likely. So for example, we can talk about a discrete random variable even if the die is loaded. So you still have six cases on a loaded die, it's just that maybe the six is a bit loaded, so it has a higher probability than the others, but it's still a discrete random variable. So now we come to one of the really important ideas of random variables, the idea of distributions. So the distribution of a random variable X is the collection of all the probabilities where elements of X exist in the sample space for all sets of real numbers such that X is an element of the sample space is, in fact, an event. So this takes us back to the formal definitions that we had in the previous parts of the course where we talked about the sample space, S, which contains all the different kind of elements, all the different possible outcomes. And every single one of these events is essentially an element, a part, value that a random variable can take, which is part of the overall set. So to look at this in more practice, look at the simple coin toss. So the probability of the random variable being heads is equal to the probability of random variable being tails, which is equal to one-half. And of course, we can talk about the distribution of, let's say, the number of heads in 10 coin tosses. So not just a single coin toss, but 10 of them, one after another. So we have 2 to the power of 10 different outcomes, and as a result, for every single one of those outcomes, which we define as small x, the probability that random variable X capital is equal to that small x is equal to 1 over 2 to the tenth because they are all identical. They all have equal probability. Now let's suppose we need to count the number of outcomes such that X as a function of s. So we want to find the number of heads, for example, such that the number of heads, the function X of s is equal to some value x, whatever that happens to be. So the number of such outcomes is the number of subsets of size x that can be chosen from the 10 different tosses, and that's 10 over x. We talked about the binomial coefficients previously in the course. So the probability that the number of heads is equal to x is

10 over x multiplied by 1 divided by 2 to the power of 10. That's the probability of an individual outcome for values of x from 0 to 10 inclusive. So the important part of this example is this idea of a probability function. So given that you have a random variable X, which has a discrete distribution, we talk about something called a probability function. And that is basically a function such that for every real number x, there is an f of x, which gives you the probability that a random variable is, in fact, equal to this x. So for example, if we talk about, let's say we talk about a die roll. So in this case, f of x is just a fixed value. It's one-sixth for x having the values 1, 2, 3, 4, 5, and 6, and it's 0 otherwise. So you're going to see this clause of 0 otherwise in many of these discrete cases, because remember, the probabilities are only available for the possibilities which the element actually gives you. So the die has six sides, and for each of those sides, you get a probability of one-sixth. But, the probability of rolling a 0 on a 6-sided die, or the probability of rolling 55 on a 6-sided die is 0, and that's what this definition of a function actually gives us. This is also sometimes known as a probability mass function. In actual fact, you're going to see this kind of physics kind of language, because later on we're going to talk about probability density. And so with relation to that, this is called by some people as the probability mass function. I'm just going to refer to it as the p robability function, or pf for short. And it's really up to you if you want to use this physics-related terminology or not.

Discrete Uniform Distribution

Okay, so let's take a look at an actual distribution. And first of all, we're going to look at the simplest discrete distribution that there is, the uniform distribution of integers. So let's suppose you have a lottery machine. So a lottery machine has these round balls corresponding to the different lottery numbers, and somebody goes up to the machine and they pull out the numbers. So there is a finite set of machine numbers in the lottery. So for example, in the UK Lottery there is, I think, 49 numbers, so you have a finite set. Now each of these balls, if the machine is not cheating, each of these balls is equally likely to be drawn, so the probability of drawing ball number 33 is equal to 1/49 on the first draw. Of course, after the first draw, it depends on whether you replace the ball or not. If you take out the ball, then the probability of something else would be modified. So for example, the probability of X = 33 on the second draw would be 0 because you've taken that ball out. But the probability of X being equal to 1, for example, would be 1/48 because, well, 1 ball has been removed, but the other probabilities are equal. Anyways, we talk about the uniform distribution on k integers as having the probability 1 over k for each integer. Fairly obvious stuff. If all of them have equal probability, then you need this stuff to end up to 1 if you add up all the distinct probabilities, so that's why you have 1 over k, and then you multiply it

by k elements and you end up with exactly 1, which is what you want the final total probability of the entire sample space to be. So to generalize this, given a random integer from a to b inclusive such that a is less than b, we now have b - a + 1 different possible value. So the probability function, or the function which assigns the probability of getting a particular number in this case, is 1 over b - a + 1, provided that x is somewhere inside that region from a to b, so it's one of those numbers from a to b, and it's 0 otherwise.

Binomial Distribution

Now let's talk about the binomial distribution. So a binomial distribution models processes which can be in one of two possible states. So for example, you could have an item that's coming off the manufacturing line, and it's either defective or not. And it can be defective with some probability, p, for example. And what we want to find out is we want to find the probability of x items being defective in a production run of n items. So you're making let's say 100 items, and you want to find the probability that 5 of those items are defective, because that tells you a lot about how the manufacturing equipment is actually working. So we can talk about sequences of failures and successes. So you have x failures and n - x successes. Now I've kind of separated them into the left and right parts, but you could have one after another. You could have failure, failure, success, success, for example. I'm just making it a bit more clear. So if you group them into failures and successes, you'll end up a certain part of failures, let's call it x, and a certain part of successes, let's say n - x successes. And in this case, the probability of exactly x items being defective, or the probability of n - x items being non-defective, is p to the x multiplied by 1 - p to the power of n - x. Hopefully, that much is obvious from our previous definitions of probability, the fact that you simply plug in the probabilities for failures and successes into the above kind of layout, and you get this very simple expression. So now let's talk about what is the actual number of sequences of these success-failure pairs. Well, we already know that it's basically the binary coefficient, so the number of sequences of success-failure pairs, provided you have a finite set of elements, is n over x, where n is the total number of elements. So from that, we can have an expression for the actual probability function. So it follows that the probability of X having some value small x is n over x multiplied by that probability that we've looked at on a previous slide. In other words, the probability function of x can be expressed as follows. So for x having an integral value 0, 1, and all the way up to n, you have this expression, and the probability is 0 otherwise. So the distribution represented by this probability function is called the discrete binomial distribution with parameters n and p. So it's not just a distribution where you give the x and you get something out

of it. You also have to specify additional things. You have to specify these parameters, n and p, for the whole thing to actually work.

Geometric Distribution

All right, so now we're going to talk about the geometric probability distribution, which is somewhat similar to the binomial experiment with a success probability p, except that we're measuring a different thing. So remember, in the binomial experiment, what we're interested in is we're just interested in the proportion of elements that succeeded, or the proportion of elements that failed. Now here, we're measuring something different. Here, what's happening is the random variable actually represents the trial on which the first success occurs. So here is the idea, here is the illustration. We have a sample space consisting of sample points E1, E2, E3, and all the way up to En. So essentially, one possibility is that you got a success on the first trial. Another possibility is that you get a failure, and then you get a success on the second trial. Another possibility is that you get a success on the third trial, all the way up to having n - 1 failures followed by a success. And what we're interested in, of course, is we're interested in how many steps do we actually have to take before we get a success? So here's the idea. We have this random variable, X, which represents the number of trials up to and including the first success. Now what you'll notice about every single one of these events, En, it does not include any of the prior outcomes. Because what's happening is as soon as we get a success, we kind of stop executing the whole process. We get a success, we're happy with it, and so every single one of those events is kind of independent of the other events. So these trials are independent. And what we can say is that for x having value 1, 2, 3, and so on, we can formulate a probability function. So the probability of getting x - 1 failures followed by a success is just multiplication of the corresponding probability. So remember, q is 1 - p, so you effectively get q to the power of x - 1, multiplied by p. And this gives us the formal definition of the geometric probability distribution. So a random variable X has geometric probability distribution if and only if the probability function of x is q to the power of x - 1, multiplied by p, where x is a whole number. So just 1, 2, 3, and so on, p is a probability value between 0 and 1, and of course, q is, by definition, 1 - p. So let me give you an example of a way you would use a geometric distribution. So for example, suppose the probability of some engine malfunction in a 1-hour period is 0. 03. So what you want to find out is you want to find out the probability that the engine will survive 2 hours of operation. So we let X to denote the number of 1-hour intervals until the first malfunction, and then we have the following. So the probability of surviving 2 hours is the probability that this random variable, which specifies how many hours you've actually survived, is greater than or equal to 3. So that would be the sum of

the probabilities with y from 3 all the way to infinity. Now we know that the total probability is equal to 1. So the sum of probabilities from 1 to infinity is definitely equal to 1, because, well, that's our definition of probability, basically. So the probability that we'll survive 2 hours is 1 minus the sum from 1 to 2 of p of x. So that's 1 - p - q times p, so we can plug in the numbers here. We can plug in 0. 03 and 0. 97, and we end up with a value of 0. 9409. So that's the probability that our engine, which has a probability of malfunctioning of 0. 03 per hour, will actually survive 2 hours of operation.

Hypergeometric Distribution

So now that we looked at the geometric probability distribution, we're going to discuss the hypergeometric probability distribution. So if you consider a population of N different elements, and each one of those elements has some characteristic with two possible different states. So for example, let's suppose that you have a bag of balls, and the color of this ball can be, for example, red or blue. So you have r total elements, which are red, and b, which is equal to N - r, total elements, which are blue. Now you take out n elements from this bag. So you sample n elements out of the overall population. And what you're interested in is the random variable X, which is the number of elements in successful cases. So in our case, success might be, for example, red balls. So you have some red balls, you have some blue balls. You know they're a number, and they're all in a bag. You pull out a certain number, and you want to find out the probability that, for example, all the balls are red, for instance. So what happens here is that this random variable X follows a hypergeometric distribution. Now it's derivation, derivation of the actual formula is a bit too complicated, so I'm just going to show you the final definition, so to speak. So here it is. As you can see, it's a bit scary. Lots of binomial stuff in here. So essentially, a random variable X follows a hypergeometric distribution if its probability function is what you see on the screen. As you can see, we have lots of parameters here to plug into this. So we have capital N being the population size, so the overall number of balls that you've got in the bag. R is the number of success states. In our case, it's the number of red balls, for example. N is the number of draws, so how many elements you actually pull out of the bag. Obviously without replacement, so you don't put the ball back in. You sort of take it out, and that's pretty much it. And x is the number of observed successes. So x is the number of elements that you actually expect to see. So you can calculate, for example, the probability that even though you pull out three balls, you want to find out the probability that at least two of them are red. So that's another sort of corollary, I guess, or another extension to the way that this formula can be used. So the hypergeometric distribution, let's take a look at an example. Let's suppose you have a factory which has 10 machines, and 4 of

those machines are defective. So if we pick five machines at random, what's the probability that none of them are actually defective? Well, first of all, we know that if four of them are defective, then six of them are non-defective. So we have some values to plug in. We have N = 10, that's the total number of the machines. We have r = 6, that's the number of non-defective machines. We have n = 5, which relates to how many elements you actually want to pick. And we have x = 5, that's how many you draw. And so the probability can be calculated just by plugging in the values inside the probability function. And you get 1/42, or about 0. 00238, so that's your answer.

Continuous Distributions

So now that we've spent considerable time discussing the discrete distributions, we're now moving on to continuous distribution. So what is this all about? Well, under a continuous distribution, we actually assign a probability of 0 to individual values. So P of X = x is 0 for each value of x. So for example, the probability that a person is exactly 2 meters tall is 0. And you can replace that 2 meters with some other value, and the probability will still be 0. So what this implies in practice is that a probability function, which assigns a probability to a particular case, makes absolutely no sense because it's kind of 0 everywhere. But we can talk about probabilities for continuous distributions. And in order to do this, we're going to introduce a new mathematical apparatus. So what we can talk about is we can talk about the probability that the random variable falls between some two values, between values a and b. So in this case, what we can say is that this random variable falls between some bounds with a sudden probability. So for example, if you think about newborn babies, for example, you're going to find a high probability that most of the newborn babies are between let's say 50 and 54 cm, and then you have the outliers. So you have the prematurely born babies, for example, and they would be smaller. Or you have babies which are huge, and they would be larger. And this way, it's easier to talk about the probabilities of getting a measurement on a newborn baby. So we're going to start somewhere, and we're going to start by introducing something called a distribution function. So given some parameter x, we define a cumulative distribution function as the probability that the random variable is less than or equal to a particular value. So with this definition of distribution functions, we can actually introduce further bits of mathematical apparatus to talk about the probability of a value falling in some range or other. So let's take a look at an example of a cumulative distribution function. So let's suppose that you have a random variable which has a binomial distribution with parameters of n = 2 and p - 1/2. So this is a production line which produces 2 items, and each item has a 50% chance of being faulty. So what we want to be able to do is we want to find the probability function, first of all, and then find the cumulative distribution function.

So the probability function, and you will notice that I'm still working on a discrete distribution here, the probability function for this distribution is as shown on the screen, essentially. We've already talked about the probability function for a binomial distribution, so no magic here. And we can plug in the discrete values of x. So you can have x being equal to 0, 1, or 2, that's the maximum number of items, and you get the following probability value. So the probability of 0 items is one-quarter. The probability of 1 item is one-half. And the probability of 2 items is one-quarter. So remember, that's the probability, for example, of the items being faulty when you produce two items on the conveyor line. Okay, so with this value, what we can do is we can move on from the probability function to the cumulative distribution function. So essentially, let's actually plot this. So we're going to plot the cdf, the probability that the random variable, the number of items which are faulty, is less than or equal to some value. So let's remind ourselves of the values of the probability function. So the probability that the conveyor belt that makes two items has no faulty items is one-quarter, The probability that there is one faulty item is one-half, and the probability of two faulty items is once again one-quarter. So for each of the values of capital F of x, that's our cumulative distribution function, we add up all the different probabilities where a is less than or equal to x. So we add up all the cases that we've met already. And this is why it's called cumulative, because we accumulate the probabilities. So we get the following values. So for 0, that's the probability that the number of faulty items is less than 0, obviously, that's impossible. So you can't have the number of faulty items less than 0. So the probability is 0 here. Now for x being between 0 up to, but not including 1, the probability is one-quarter. Now as you reach the value of 1, you include the second probabilities. So we move from one-quarter to three-quarters. And then as you reach, go above 2, or reach 2 and above, the probability is 1 because that basically includes all the possible cases. So if you were to plot the distribution function, it's going to look something like this. It's just a bunch of lines where up until 0 you have the value of 0, because there is nothing there to generate any probabilities. Up until 1, you have a value of one-quarter. Up until 2, you have a value of three-quarters. And then you have the value of 1 from 2 all the way up to infinity. You have up to however many other cases you might want to consider. Okay, so with this setup, we can talk about some properties of a distribution function. So the distribution function at negative infinity is equal to 0. Essentially, this is the idea that towards the very left of the scale where you don't have any items to consider, the probability is on the lower bound, which is 0. Remember, we have this agreement that probabilities stretch from 0 to 1, and of course, F of infinity. So on the rightmost side of the graph, when you've included every single case, every single possible probability, then the value is equal to 1. Now we also require that the cumulative distribution function is a nondecreasing function. If it were a decreasing function at any point, that would imply that you can have negative probability. And

remember, we have this assumption that a probability of any particular event is from 0 to 1. So it cannot be negative by definition, because that's the way we've defined probability. Now there are ways of defining probabilities differently, but we're not going to go into that in this particular course. So one thing we can say about random variables is we can say that a random variable X is continuous if its cdf is continuous for the entire range of x. So that's an alternative way of defining a continuous random variable. All right. So now that you know about the cumulative distribution function, we're going to take a look at something which is perhaps a bit more important, and that is the probability density function. So here is the definition. If capital F of x is the distribution function for a continuous random variable, then we define f of x, notice that's a small f of x, as the derivative of the distribution function. And this derivative that we acquire is called the probability density function on the random variable X. So what can we say about this? Well, we can say that f of x is greater than or equal to 0 for all x. That is hopefully obvious. But another thing that we want to require is we want the integral from minus infinity to infinity of f of x being equal to 1. In fact, that's a requirement for a function to actually be a probability density function. Meaning that if you have a function and you take its integral from minus infinity to infinity, and you don't get a 1, it doesn't represent a probability density function. It's as simple as that. So now let's take a look at how you would actually calculate any kind of probability values using a probability density function. So let's suppose we have the probability function f of x. Then what we can say is that the probability that this random variable actually lies between points a and b is in fact the integral of the probability density function from point a to point b. So let's suppose we're given the following probability density function. So f of x is equal to 1 over 8 x, where x goes from 0 to 4, and it's 0 otherwise. By the way, you'll notice that if you were to integrate this from 0 to 4, that's where x is actually defined, if you were to integrate this, you would get a 1. You didn't need the integral to be equal to 1 on the entire range of defined values of x for this to actually count as a probability density function. So now you can actually calculate the probability, for example, that the random variable is between 1 and 2. So that is basically the integral from 1 to 2 of 1 over 8 x dx, and that is equal to 3/16, for example. Or you can calculate, for example, the probability that X is greater than 2. That is the integral from 2 to 4. Effectively, you would have an integral from 2 to infinity, but we know that x is defined from 0 to 4, so there is no point going beyond 4. So that is the integral from 2 to 4 of 1 over 8 dx, which is equal to three-quarters.

Continuous Uniform Distribution

So now that we know about the cumulative distribution function and the probability density function, let's take a look at a uniform distribution once again. Now we already have the uniform

distribution in the discrete case. So where you roll a die, for example, and all of the probabilities are equally likely. But here, let's consider a case where you have a continuous distribution of the variable. So for example, let's imagine that you have a train, and the train always arrives between let's say 6:30 and 6:40. Now the probability that this train will arrive in any subinterval from 6:30 to 6:40 is proportional to the length of the subinterval. So let's suppose we use the variable, the name X, to denote the amount of time that a person has to wait for a train if they arrive at a station exactly at 6:30. Then we can say that X has a continuous uniform probability distribution. So here is the formal definition of the uniform probability distribution. So if a is less than b, then a random variable X is said to have a continuous uniform probability distribution on the interval from a to b if and only if the density function of X is defined as follows. So essentially, you can see that f of x is 1 over b - a when x is on the interval from a to b, and it's 0 otherwise. And the constants a and b are the parameters of this particular density function. So let's take a look at an example of the application of the uniform probability distribution. So let's suppose that trains arrive within a 30-minute period. What is the probability that the train will arrive in the last 5 minutes of that interval? Now you can already kind of figure this out in your head, but we're going to apply the actual density function here. So we have a uniform distribution with parameters a = 0 and b = 30. So the probability that X is between 25 and 30 is the integral from 25 to 30 of 1 over 30 dx, which is equal to one-sixth, also a very intuitive solution here.

Normal Distribution

Next up, we're going to talk about a very popular distribution. And this is the normal probability distribution. So let's take a look at the definition first. So a random variable X has a normal probability distribution if and only if for a sigma greater than 0, and mu being somewhere between minus infinity and infinity, the density function is as follows. So you have 1 over sigma root 2 pi multiplied by the exponent of x - mu squared divided by 2 sigma squared. Seems a bit complicated. Now this normal density function has two parameters, mu and sigma. And a distribution with mu = 0 and sigma = 1 is called the standard normal distribution. Now I've actually pushed back the discussion on things like the mean and the standard deviation until the next module of this course. But essentially, the mu here is the mean, and the sigma is the standard deviation. You're going to encounter them a bit later in the course, but I just wanted to mention it here. So what can we say about the normal distribution? Well, first of all, as I mentioned, the standard normal distribution with mean of 0 and standard deviation of 1 actually simplifies the formula a little bit. Because if you plug in all the numbers, you just get f of x = 1 over square root 2 pi, e to the -x squared over 2. As you may have guessed, the indefinite integral of e to the -x

squared over 2, so the integral from minus infinity to infinity of e to the -x squared over 2 is, in fact, equal to square root of 2 pi, because you need the integral of the density function to be equal to 1. However, here is the interesting part. So to find out the probability of x being between some bounds a and b, we would need to evaluate the definite integral, the integral going from a to b of e to the -x squared over 2. And yes, there is a 1 over root 2 pi there, which I omitted, because that's not where the problem is. The problem is there is no closed-form solution for this particular integral. So in order to evaluate this integral, you would need numeric integration techniques. So if we're talking about the R language, for example, you do get those techniques. You do get the functions for evaluating it. So we have two functions. We have pnorm, which gives you the probability that the random variable is less than or equal to some value x. So it takes three arguments, x, mu, and sigma, and it gives you the probability of the x, which is distributed normally with mu and sigma is less than or equal to x. Or there is also the quartile function, the pth quartile. So you call qnorm. So notice the first one has a p in front, pnorm. The second one is qnorm. So you call qnorm with p, mu, and sigma, and here the idea is as follows. It gives you the value of x such that the probability of the random variable being less than or equal to p is equal to x. So you have these two possibilities for calculating the values inside the distribution. And other distributions, as we'll see later on, also have this pair of functions for figuring out either just the probability or the pth quartile. So let's take a look at an example of using the normal distribution. So let's suppose we have a bunch of test scores, and we know that test scores are normally distributed with a mean of 75 and a standard deviation of 10. So the question is what fraction of the scores lie between 80 and 90, for example? Before we got all sophisticated with computers and calculators and whatnot, we calculated things using tables. And if you wanted to calculate the answer to this question using tables, you would first of all perform a transform of the variables. Basically, you can transform this particular distribution with a particular mean and standard deviation to a standard distribution by using this transform. So you take the variable, you subtract the mean, and you divide it by the standard deviation. And this gives you new values for the bounds, so to speak. So it gives you a transformed value for the first and second bounds of the calculation that you want to do. So now we have 0. 5 and 1. 5 as the bounds, and you would look up these values inside a table of the standard distribution, and you would subtract one from another. Now of course, since we have languages such as R, we can just do it using the functions that are built in. So we say pnorm 90, 75, 10. So as you can see, the first value here is the test score on the upper bound, the second is the mean, and the third is the standard deviation. So we say pnorm of the right side minus pnorm of the left side. And that way, we get our final value, which is 0. 24173. So about 1/4 of the overall test scores lie between 80 and 90. Now normal distribution is actually used extensively in natural and social sciences. It's used literally all over the

place, so I won't bother listing every single application of it. The one that's particularly interesting to me is Brownian motion, which is used not just in physics where it originated, but also in areas such as mathematical finance where it's used to model the price of stocks, for example.

Gamma Distribution

Now we're going to take a look at an even more sophisticated probability distribution called the gamma probability distribution. So a random variable X has the gamma probability distribution with positive parameters alpha and beta if and only if the density function is presented as below. So essentially, you have a fraction of where on the top you have x to the alpha - 1 multiplied by e to the -x over beta. But it's the bottom part that's interesting, because on the bottom, we have the gamma function, gamma of alpha. So the gamma function is a special integral. It's an integral from 0 to infinity of x to the alpha -1 multiplied by e to the x, and surprise, surprise, this integral does not have an analytic solution. Once again, this is an integral which has its own special function simply because we cannot give you an expression in terms of x which is a closed-form solution. So let's actually discuss the gamma function. So essentially, a gamma function is a special function. So it represents an integral that I've just talked about, and it's called, yes, the gamma function. So the gamma of 1 is fairly easy to calculate, because it's the integral from 0 to infinity of e to the -x, which is equal to 1. And then what you would do if you wanted to evaluate this integral is you would do integration by parts, and you wouldn't get anywhere, but you would get to the following interesting relation. So the gamma of alpha is equal to alpha - 1 multiplied by the gamma of alpha - 1. So looks familiar, doesn't it? Looks a lot like the factorial. And an actual fact for natural numbers, the gamma of a natural number n is equal to n - 1 factorial. Now if you look at the plot of the gamma function, it doesn't really look like much, and it doesn't really help us in any way to sort of visualize or figure out what this function is all about. Now let's actually go ahead and plot the probability density function of the gamma distribution. So here is the function, f of x. So as you can see, we have to specify the parameters alpha and beta for this to actually be a valid plug. So what we're going to do is we're going to assign alpha a bunch of different values like 2, 4, 6, and 8, and we're going to fix the beta as the value of 1. And here is what we get. So this graph hopefully illustrates why alpha is called the shape parameter and beta is called the scale parameter. So as I'm varying the alpha, you'll notice that the shape of the graph fundamentally changes from a sort of a vertical shape to a more flattened-out shape. The gamma distribution is used in plenty of places. It's used in insurance claims. It's used in modeling the amounts of rainfall, surprise, surprise. It's used in measuring signal propagation in wireless communications. It's also used in measuring signal propagation in the brain, so in neuroscience.

And it's also used in advanced statistical applications like multi-level Poisson regression models, for example.

Beta Distribution

All right, now we're going to look at the beta probability distribution, even more complicated than the gamma. So a random variable X is said to have a beta probability distribution with parameters alpha and beta, which are both greater than 0 if and only if the density function is as shown below. So once again, we have a fraction. And on the bottom of that fraction, we once again have a special integral. This time round, it's an integral from 0 to 1 of x to the alpha - 1 multiplied by y - x to the beta - 1. And surprise, surprise, this integral does not have an analytic solution, but it can be presented as a relationship of gamma functions. So it actually, beta of alpha and beta, so the beta function of parameters alpha and beta is equal to gamma of alpha multiplied by gamma of beta, divided by the gamma function of alpha plus beta. So fairly complicated stuff. So the cumulative density function, distribution function rather, for the beta random variable is called the incomplete beta function, sometimes represented with capital I subscript x. So that's the integral from 0 to some value x of that thing that we just looked at, that entire fraction. So when alpha and beta are positive integers, we can once again do integration by parts, and we can get a sum as the output. And this sum also includes a binomial term in there, so binomial term, n. And this n is a newly introduced variable. So n is actually alpha + beta - 1. And if you look at what we're actually getting, so the binomial term here, you can see that the sum that we're getting is the sum of probabilities associated with a binomial random variable with the n value representing alpha + beta - 1, and the p value, the probability value of success, of beta being directly related to the x variable in our example of the beta distribution. So let's actually do a plot of the density function by fixing either alpha or beta. So in the first case, what we're going to do is we'll specify alpha having the value of 2, and beta having values of 1/4, 2, and 4. And in the second case, we'll fix the beta, so we'll say beta is going to be 1. 5. And we'll change alpha to 1/4, 2, and 4 as well. So here is the first plot. That's where the alpha value is fixed. And you can see that there is very little in terms of kind of predictability when it comes to modifying the beta value. And similarly, down below you can see what happens when a beta is fixed. So depending on the alpha, you have either the whole thing skewed towards the left, or kind of trying to go towards the right like a wave. Feel free to experiment with other values, and you'll get different variations on this plot. So once again, for calculating the beta distribution, you have pbeta, which yields you the probability that X is less than or equal to some particular value. You'll notice, though, that the third parameter here is not beta itself. Instead, it's 1 over beta. So when you actually feed in the values to get your

beta distribution parameters, you actually have to specify 1 divided by beta. And the same goes for qbeta, which actually gives you the pth quartile function. So once again, you feed in p and it gives you the probability that a random variable gives you the value, in actual fact, that the probability of the random variable being less than or equal to this value is equal to p. So those are the functions for calculating the values in a beta distribution. And I have to say that the beta distribution is a lot less practical in the sense that it's primarily used within statistics itself, rather than any kind of natural application like physics, for example.

Summary

All right, so let's try to summarize what we've learned in this module of the course. So we looked at discrete distributions, and the idea of having the discrete distribution characterized by a probability function. And we looked at some discrete distributions such as the uniform distribution, binomial distribution, geometric, and hypergeometric distributions. And then we discussed continuous distributions, which cannot be characterized by a probability function, but they can be represented by a probability density function, which happens to be the derivative of the cdf, the cumulative distribution function. And we looked at some continuous distributions such as the uniform distribution, the normal distribution, as well as the gamma and beta distributions.

Introducing the Concept of Expectation

Overview

Hello, and welcome to this module of the statistics course. In this module, we're going to discuss the concept of expectation. So the goal of this course might appear to be simple. I want you to understand what expectation is in the mathematical sense, but we're also going to take a look at quite a few other statistical measures which rely on your understanding of the concept of expectation. So here's what we're going to see as part of this module. First of all, we'll discuss what expectation is and what it's used for. We'll discuss the associated idea of the mean or the average of a data set, and then we're going to talk about a few other things. We'll discuss of Law of the Unconscious Statistician, we're going to discuss the idea of variance, we'll take a look at

moments and the idea of a moment generating function, we'll take a look at joint distributions, and we'll finish off the discussion with a look at covariance and correlation.

Expectation

All right, so let's get started, and as I promised, we're going to start by discussing this idea of expectation. What is expectation all about and why should we care about it? Well let's imagine that you want to have an estimate for the height of a newborn baby, so you get newborn baby and how tall do you expect that baby to be? Would you accept, for example, if I told you that the newborn baby is 1 meter long or 10 meters long? Of course you wouldn't because it's extremely unlikely that anyone could produce such a baby, but we have to do this in a more formal fashion, so we don't really know the baby's height in advance because it's a random variable, but what we can do is we can try to get an estimate for the kind of, the kind of range is the kind of sizes that this variable can actually take, so here, we can take a bunch of newborn babies. Let's say n different newborn babies, and obviously the more newborn babies you take, the better, the more accurate your result will be, and we can average out the result. So this is something that you should remember from your math course in school, so to calculate the average, you simply add up all the babies' heights and divide them by n. So if you took 100 babies, you would add up their heights and you would divide them by 100, for example. Okay, so this would give you what we call the mean, or the average baby height. So this would be a value which is kind of like the central value or the most expected value a newborn baby would actually be in terms of their height. So given, if we now formalize this, given a random variable, capital X, which represents the baby's height, then we say that this X bar value that we've just calculated is the expected value of that random variable. Now you can see that there are several different terms here which mean pretty much the same thing. So we have mean and average and expected value, or we can also talk about the expectation of a random variable X, and they all mean the same thing. They just mean the average, basically, so in the case of babies, the average is about 51 cm, or 20 inches, and what this means is that as you're getting a new baby, a newborn baby, it's very likely that its height will be closer to, let's say, 51 cm than 50 cm or 55 cm. That's what the average actually gives us. Now one thing which this experiment does not give us is it does not really give us any kind of probabilities. There is no assumption that a baby which is 1 meter long is improbable, even though in the real world we know that it's pretty much impossible. So, given this result, we now want to kind of expand this idea of expectation, and we want the idea of expectation to cover not models where you simply pick a few elements and average them out, but we want to apply the idea of expectation to models where we already know certain probabilities. So here it is. So the

expectation of a random variable, which we also call the expected value or the mean or the average, so the expectation of a random variable is a weighted average of all the possible values weighted by their respective probabilities. So if you already have probabilities for the different kind of values that a random variable can take, then you simply take a sum of their respective products. So it's value times probability plus another value times its probability, and so on and so forth. Let's actually take a look at an example. Okay, so here's a very simple example of calculating an expectation. Let's suppose that you're modeling stock prices, and you know that a stock price can increase by 1 point with a probability of 0. 6. It can stay in place with a probability of 0. 1, and it can go down by 1 point with a probability of 0. 3. So now the question is if X represents the change in stock price, what is the expected value of X? By the way, let's make a pause here and discuss the notation. So, the notation I'm using here is capital letter E, and that's what everyone uses, every single book uses capital E to model expectation, and then the random variable. Sometimes it's put in round brackets, sometimes it's put in square brackets. I've seen curly braces as well. By the way, I'm going to be using square brackets here, but don't be surprised if you see round brackets, for example, it really depends on the author of the book or article that you're reading. So this is the notation that I'm going to use, so the expectation of a random variable X is written as capital E, and then that random variable in square brackets. So this is what we want to find out, and using the definition that we had on the previous slide, here is the actual calculation, so we take a sum of the product of the values themselves by their respective probabilities. So given that the probability of growing up is 0. 6 and the change is 1 point, so the stock price is going to increase by 1 point, we'll multiply 1 by 0. 6. Now, if the stock stays in place, then the change is 0, and the probability of 0. 1, but 0 times 0. 1 is obviously 0. And the last one, I've sort of taken out the - 1 just to be clear about it, so the probability is 0. 3, but the change in stock price is actually, well, since it's going down, it's -1. So you have 0. 6 - 0. 3, and the result is 0. 3. So in other words, on average, you would expect a stock price to go up by 0. 3 points.

Mean

So here is a formal definition of the mean, which is also the expected value and the expectation and the average, all of those things at the same time. So for a bounded discrete random variable X with a probability function f, the expectation of X is as follows. So it's the sum of the product of X by its probability function for the all the possible values of X. So essentially what you're doing is you're taking each of the possible values of the random variable, capital X, and you're sticking it into this product, so it's the variable itself times its probability function, and then you add them up together, and that's what actually gives you the mean. So let's take a look at an example.

We're going to take a look at the mean of a Bernoulli random variable. So a random variable X has a Bernoulli distribution when X can only take values 0 or 1 with a sudden probability of having the value of 1 being equal to P, and likewise the probability of that variable being 0 is 1 - P. So what we can do is we can try to calculate the mean of that, or the expected value. So this is once again a sum of the product of the different values, and we have value 0 and value 1 multiplied by their respective probabilities, and of course, for the value of 1, the probability is p and for the value of 0 the probability is 1 - p, so you get 0 times 1 - p, which is obviously 0, plus 1 times p, which is equal to p. So essentially, what this is telling us is that the probability of a Bernoulli distribution taking on a value, on average, so the averaged value, if you took also these values and you averaged them out, the probability of that variable getting close to p is highest compared to all of the other possibilities. So most of the variables are going to be closer to p than they are to somewhere else. You'll notice that, funnily enough, the Bernoulli distribution cannot take on the value of p generally, so let's imagine p is 0. 5. You cannot have a value of 0. 5 by itself. You can only have a value of 0 or a value of 1, but p is quite simply a model. It's quite simply a central value that does not give you a concrete point in the sample space, but it gives you a kind of indication as to where most of the values are actually going to tend towards. The mean of a random variable does not always exist, so we need to discuss this idea of the existence of the mean. So given a random variable X with the probability function f, we can consider, for example, the positive and negative sums of that variable times that function. So we can sort of split the calculation into taking all the positive values of x and all the negative values of x, and here, the idea is that if at least one of these is finite, then the expectation is said to exist, and it's therefore defined as the sum of that variable times its probability function as we've seen before for all values of X. However, if both of the sums are infinite, then that means that the mean does not exist, so you're going to encounter distributions like this, and let's take a look at one of them. So here, let's have X be a random variable with the following probability function. So this is a specifically chosen probability function, 1 over 2 modulus of x times modulus of x plus 1 for non-0 whole numbers of x, and then what we can do is we can simply check the sums on these sort of positive and negative side of things. So we can calculate the sum from -1 to minus infinity of x times 1 divided by 2 modulus of x modulus of x plus 1, so that is the negative side, and if we calculate that, we get negative infinity, and similarly we can calculate the sum from 1 to infinity of x times 1 divided by that same thing once again, and this gives us a positive infinity, and as a consequence of that, what we can conclude is that since both sums are infinite, the actual expected value or the mean of the random variable X simply does not exist.

Expectation for a Continuous Distribution

Now we've talked about the expected value for a discrete distribution. That's actually a very easy thing to calculate. You simply multiply all of the different values since there's is a finite set of values, you multiply them by their respective probabilities, or probability functions, apply to those values, you add them up together and that's your expected value. That's very simple. Now let's talk about continuous distributions. So for a continuous bounded random variable, just as we've seen before, what happens is a sum actually turns into an integral and you end up with something like the following. So if random variable X is a bounded continuous random variable with a probability density function f, then the expected value is quite simply the integral across the whole space from minus infinity to infinity of that variable times the probability density function. So last time around it was the probability mass function that was applied to the variable, here it's the probability density, and this once again gives you something that we call the expected value, or the mean. So let's take a look at an example of a continuous expectation. So an expectation calculated for a continuous random variable, so let's suppose that we have some machinery, some machine that is working and it has a part which fails with the probability density function shown here. So it fails with the probability of 2 x, given that x is somewhere between 0 and 1. So with this setup, if we want to calculate the expected value of that variable, we would integrate, and obviously in the real world you would integrate from minus infinity to infinity, however, we know where x is bounded, so we know that in this particular case, x only exists between 0 and 1, and it's 0 in all other locations so we can take that integral from 0 to 1, and under the integral, we have x multiplied by f of x. So x times 2 of x, which means you have 2 x squared, you integrate that from 1 to 0, and you get two-thirds, so the expected value is two-thirds. That's when you actually expect that part to fail. Similarly to discrete distributions, you also have the idea of the existence or nonexistence of a mean for a continuous distribution. So here the idea is very simple. If either the integral from 0 to infinity or minus infinity to 0 is finite, then the mean or the expectation or the expected value of X actually exists, and it's defined as that full integral. However, if both of the integrals are infinite, if both of them give you infinity and minus infinity, then the mean does not exist. So let's take a look at an example of attempting to figure out whether the mean of something exists or not. So we're going to consider the Cauchy distribution, which has the following probability density function. It's a very specifically chosen function. You can see that there is a pi in there, and the reason why it's a pi is because, well, remember, you have to have the PDF have an integral of 1 overall, and this is done deliberately because the derivative of an arc tangent is actually that 1 over 1 plus x squared, so that explains why the pi is there. So we can try to find out whether the mean for this distribution actually exists. So we can calculate the 2

integrals, the integral from 0 to infinity, the integral from minus infinity to 0, and you can see that they're equal to infinity and minus infinity, respectively, and as a consequence, we can conclude that the mean for the Cauchy distribution, once again, does not exist.

Functions of a Random Variable

Having talked about random variables, we're now going to discuss this idea of functions of random variables. That is when you take a random variable and you try to stick it into a function, and then you are effectively generating a new random variable by applying that function. So if X is a random variable with a probability density function f, we're going to consider some real valued function r of X, so different values of X, as they are generated, are fed into some function and you generate a new set of data. So to find the expected value of the new data set, we do the following. So first of all, we have a new random variable, and here I'm calling it Y, so we define the new random variable Y to be defined as the application of the function r to the random variable X, and then of course, we determine the probability distribution of Y, so we find out what the probability distribution of Y is, and finally, we determine the expected value of Y, because we already have a function which incorporates the idea of probability distributions. Okay, so having set all of this up, we then have the following. So the expected value of r of x is the expected value of Y, which is an integral of y times g of y where, of course, g of y is what happens when you actually plug in the PDF into r of x in the first place, and this is what you get. You get the expected value, provided it exists, of course. So let's take a look at an example of calculating the expectation of a function that's applied to a random variable. So let's suppose we have a random variable X which has a probability density function of 3 x squared defined on x being between 0 and 1. Now let's suppose that we have some function r of x which is equal to 1 divided by x. So what we want is we, first of all, want to calculate the probability density function of the new variable y, the new random variable which is created by applying the function r to the random variable X. So what happens in practical terms is you take 1 over x and you stick it into the place where you have the 3 x squared, so you have 3 times 1 over x squared, which becomes 3 y to the -4, and of course, the constraint changes, so we have the constraint that x goes from 0 to 1, but now that we've had this change of variable, we now have a constraint y is greater than 1. We have our new function g of y, and what we can now do is we can calculate the expected value of that because that is the integral from 0 to infinity of y times g of y, which is y times 3 y to the -4 dy, which is equal to three-halves. So funnily enough, these manipulations that we just did are actually unnecessary, and the reason why they're not necessary is due to LOTUS, which is an

abbreviation that stands for the Law of the Unconscious Statistician, and that's what we're going to take a look at next.

Law of the Unconscious Statistician

All right, so here is the Law of the Unconscious Statistician. So if x is a random variable and r of x is some real valued function of a real variable, then if x has a discrete distribution, the expected value of r of x is the sum across all the different values of x of r of x times f of x, where f of x is, of course, the probability function. Now if x has a continuous distribution, then the idea is fairly identical in actual fact. It's the integral from minus infinity to infinity of r of x times f of x, and in this particular case, f of x is, of course, the probability density function. So let's take a look at how you can apply the LOTUS, the Law of the Unconscious Statistician, to an example that we've already seen in actual fact. So here is that example that we've seen. We've had a probability density function of 3 x squared and we had r of x being equal to 1 over X, so now we're going to apply this law and we're going to take a look at how that simplifies the calculation. So here, the expected value of Y is quite simply the integral from 0 to 1 of r of x times f of x. So you simply multiply the function that you're applying to the random variable by the probability density function of that original variable. So we have the integral from 0 to 1 of 1 over x multiplied by 3 x squared, which gives us the same result as before. One thing to note that, in general, the expected value of a function of a random variable is not equal to a function of the expected value of a random variable. There are very few cases where it is, in fact, the case, but in general, it simply isn't.

Properties of Distributions

Having gotten ourselves acquainted with the idea of an expectation of a random variable, let's take a look at certain properties of distributions, and in particular the expectation of those distributions. So if X is a random variable for which the expectation actually exists, then first of all, we can say that if X is always equal to some constant c with probability 1, then the expected value of that random variable is obviously c. Hopefully, I don't need to prove this because it's just fairly self-evident. Now if the random variable, let's call it Y, is represented as a linear function of X, so it's a times X + b, where a and b are finite constants, so we have some sort of linear function, then the expectation of that random variable is such that you apply the expectation only to the random variable under the linear function, so the expectation of Y is a times the expectation of X + b. So this is an interesting kind of thing, interesting kind of manipulation that we're going to be

doing quite a lot while we expand out the expectations of more complicated functions. And finally, there's one more property of distributions and their expected values that's worth taking note of, so if you have several random variables, X1, X2, all the way up to Xn, and they are all random variables, such that their expectation is finite, then we have the following 2 rules. So the expectation of their sum is the sum of their expectations, and similarly the expectation of their product is the product of their expectations. And once again, these formulae are going to be useful for us later on.

Variance

All right, so now we're going to discuss one of the central topics in statistics, and that is the idea of variance. Now why do we need variance and what is it all about? Well, we have a bit of a problem with the mean because unfortunately, the mean of a distribution does not completely describe that distribution. All it tells us is there is some central point out of all the possible values, but it doesn't tell us how spread out the values are around the central point, and this is what the variance essentially does. Because the variance, for example, if you take a uniform distribution from -1 to 1, and you take a standard normal distribution, then they're going to have the same mean, so the mean is the same but the distributions are actually drastically different, so the variance is yet another measure, and in this case, the variance is a measure of how spread out the data is around the mean. So let's take a look at how to actually calculate the variance. Here is the formal definition. So if X is a random variable with a finite mean mu, then the variance of X is defined as V of X. Notice we're using the capital letter V. Just as we used E for expectation, we're using V for variance, so the variance of X is the expectation of a random variable that's generated by taking our original random variable, subtracting the mean, and then squaring the whole thing. So that might seem a bit weird, but this is the definition of variance. This is how we can calculate variance of any particular random variable. Now if our random variable X has an infinite mean, or if the expectation of X does not exist, then the variance doesn't exist either. Now in addition to the mean, we have this idea of standard deviation, and standard deviation is quite simply the non-negative square root of the variance. So the notation that we use, and you can see me using the mu up above, so mu is used to represent the mean or the expected value of random variable. Sigma is used to represent the standard deviation, and sigma squared, specifically, is used to indicate the variance. So let's take a look at an example of variance. Let's suppose that we have a coin flip. So we're flipping a coin and it gives us either a 0 or a 1 with equal probability of 0. 5, so we're going to try and calculate the variance for this particular experiment. So first of all, let's calculate the mu or the expected value. That is quite simply the sum of the products of those

values together with their probability, so it's 0 times 0. 5 + 1 times 0. 5, which gives us 0. 5. So that's the expected value of a coin flip in this particular case. And then we define a new variable, let's call it Y, which is defined as X minus the mean squared, and then the variance of the variable X is the expected value of that new variable Y. So we compute the probability function of Y. This is the table that shows us x, y, as well as f of y, and then we take these values and we actually plug them into the calculation. So the variance of X is the expected value of Y, which is one-half times a quarter plus a quarter, which is equal to one-quarter in the end. So basically, what we say is that the value of the random variable, which represents a coin flip, has a variance of 0. 25, and we can calculate the standard deviation as the square root of that and we get 0. 5. There is an alternative way of calculating variance, and it comes from the expansion of those terms under the expectation, so let's take a look at how that works. So here we have the variance of X defined as the expectation of X - mu squared. Now interestingly enough, what we can do is we can multiply out X - mu squared, so we get X squared - 2 mu X + mu squared, and then we can use some of the rules that we've looked at when we discussed expectations. We can break this apart. So we can break this sum into the expectation of X squared -2 mu times the expectation of X + mu squared. Now remember the expectation of X is the same as mu by definition, so you have -2 mu squared + mu squared on the right. So this leaves us with the expectation of X squared - mu squared. So to sum things up, the variance of the random variable X is the expectation of that variable squared minus the expectation of that variable, and then the whole thing squared. Variance also has a very nice additive property in the sense that if you have a bunch of different random variables and they have to be independent, obviously, different independent random variables, X1, X2, all the way up to Xn, and you also have some constants like a1, a2, a whatever, as well as a constant b, then the variance of a1X1 + a2X2, a3X3, all the way up to anXn + b, let's not forget that constant, is equal to, and then you can factor out those variables, a1, a2, and all the way, but they're going to be factored out squared. So you're going to have that variance being equal to a1 squared, V X1 + a2 squared V X2, all the way up to a n squared V xn, and notice that the b disappears completely. The b does not really contribute to the variance because all the variable b does is it kind of shifts the distribution left and right, but since variance measures how spread out the distribution is, the shift itself doesn't really matter.

Moments and the Moment Generating Function

We're now going to introduce a somewhat more complicated topic of moments. So what are moments and what is a moment generating function? That's the idea here. So given a random variable X and some value, some constant k, which is a natural number, so 1, 2, 3, all the way up to

however high you're prepared to go, the expectation of X to the power k is the kth moment of the random variable X. So this is the formal definition. Now the mean, the expected value of X, is obviously the first moment of X, so if you take k being equal to 1, then E, the expectation of X to the power of 1 is simply the expectation of X, which is the first moment of X. So the mean happens to be the first moment of X, and the kth moment exists if the following condition holds. So we need the expectation of the modulus of the variable to the power of k to be less than infinity. So we're using the modulus here because otherwise, we would have to write it using infinity and minus infinity, so here it's kind of simplified. So we want the expectation of the absolute value of the variable to the power of k to be less than infinity. In addition to the ordinary moment, there is also something called the central moment. Now remember when I talked about variance, I said that variance is a measure around the mean. So now we're going to see this kind of generalized. So an nth central moment, mu n, notice the ordinary mu is for the mean, but mu n is basically the nth central moment around the mean, so this is the moment of a probability distribution about the mean. So essentially, it's described as follows. It's essentially the expectation of X minus the expectation of X to the power of n, and you'll notice that we have the power of 2 when we talked about the variance. So here, it's defined as the integral from minus infinity to infinity of X - mu to the power of n times the probability density function. So we can plug in a few values. For example, we can plug in the value of 0, and if you plug in the value of 0, you're going to get, obviously, x - mu to the power of 0, which is equal to 1, so you'll have the integral from minus infinity to infinity of f of x, which is equal to 1 by definition because that's what we require of any valid probability density function. So the first, the 0th rather, central moment around the mean, since we're starting with 0, is equal to 1. Now if you plug in the value of 1, you're going to get a 0, so you have x - mu to the power of 1 times f of x. Now if you look at just the expansion of X minus the expected value of X, you would have, if you split this up, if you split up the sum, X minus expected value of X, you would have expected value of X minus the expected value of the expected value of X. Now there is something called the tower property which tells us the expected value of an expected value of a variable is just the expected value of that variable, and so you subtract the expected value of X minus the expected value of X, and you get 0. So the first central moment of a random variable X happens to be 0, and of course, the second central moment, as I mentioned, is the variance. That's why you plug in the value 2 for the value of n and you get the canonical representation of the variance. So having talked about moments, we can now talk about something called the moment generating function, and the moment generating function is a function which completely describes a particular distribution. So given some random variable X, for each t, for each real number t, the moment generating function of X is defined as the expectation of the exponent of t times that random variable X.

Now I know it might seem cryptic and a bit pointless, but here is one practical use of the moment generating function. So given n as a whole number, the nth moment of X actually happens to be the nth derivative of the moment generating function at a point when t is equal to 0, so the expected value of X to the power of n is actually the nth derivative of the moment generating function psi at a point where t is equal to 0, and this is the result that we're actually going to use, and it should be a lot of fun because we've never had such a shortcut before. So now we're going to take a look at an example where we're going to calculate the moment generating function, and then using the moment generating function, we're going to calculate the variance. So let's suppose we have a function X, which has a probability density function of e to the minus x for values of x greater than 0. So first of all, we can find the moment generating function. So that is the expected value of e to the power tX, and that's the integral from 0 to infinity of e to the power of tX times e to the power of -x, and so you can group the t and the -1 together, so you get e to the power of t - 1 multiplied by x. Now as you can see, if you plug in a value of t that's too high, this function is going to diverge and we're not going to get a finite integral, so we're going to talk about the moment generating function only for values of t which has less than 1. So in this case, the psi of t, the moment generating function, is 1 divided by 1 - t with a constraint that t is explicitly less than 1. So now that we found the moment generating function, we can calculate the variance easily from the moment generating function. So we found the moment generating function psi of t, which is equal to 1 divided by 1 - t, and then we can calculate its first and second derivative, so we simply differentiate psi of t, and we get 1 divided by 1 - t squared, and for the second derivative we get 2 divided by 1 - t cubed. Now what we can do is we can remind ourselves of the definition of the variance and we can easily calculate the variance. So remember, the variance is the expectation of X squared minus the expectation of X squared. So this, using the definition of the moment generating function as we've seen before, is the second derivative at t = 0 minus the first derivative squared. So we plug in a value of t = 0, we get a 2, a value of 2 for the second derivative, we'll get a value of 1 for the first derivative, so we get 2 - 1, which is equal to 1. So therefore, the variance of the random variable X is equal to 1.

Means and Variance of Some Distributions

I wanted to talk briefly about the means and variances of some of the distributions that we've already seen as part of a previous module in this course. So we're going to begin by looking at the discrete distributions. That's what you're looking at on this particular slide. So let's start with the uniform distribution, which has a probability mass function of 1 divided by b - a + 1. So essentially, the uniform distribution is where you pick a number at random from a to b inclusively.

So the mean of that is obviously a + b divided by 2 because that's the entire range, and if you pick values from that range, you're going to be closer to the middle than any particular edge. Now the variance is a bit peculiar because it has that +1 and -1. If we look at the variance for a continuous distribution, that's something that I have in the next slide, you're going to see those +1 and -1 actually disappear. Now the reason why they're here is because this distribution actually has endpoints, so when you're picking a number from a to b inclusive, you have that endpoint. So if you do b - a, you've effectively thrown out one of the values, which is why the +1 and -1 are there. Now let's take a look at the binomial distribution. So we've looked at that already. It has an average of n times p, so given n different elements, you have n times the probability as the mean, and np times 1 - p as the variance. The geometric distribution also has very neat mean and variance, 1 - p over p and 1 - p over p squared. The hypergeometric has a good-looking mean, so it's n times r divided by N, but the variance is really complicated. Now let's take a look at the continuous distribution. So here is that uniform distribution that I mentioned. Its Pdf is 1 divided by b - a. The mean is obviously the same as with the discrete, so a + b divided by 2, and the variance doesn't have all of that +1 and -1 business because remember, the probability of any particular value in the continuous distribution is 0, so we don't care about the endpoints anymore, which is why the variance here is simply b - a squared divided by 12, and if you're wondering why the 12 is there, then you can calculate this yourself using the definition of the variance that has been provided as part of this course. And then we looked at the normal distribution, and the normal distribution, in its very definition, already gives you hints as to the mean and the variance because it has the mu and the sigma in there, so the mean is obviously mu and the variance is sigma squared. In other words, the standard deviation is sigma, and we discussed this idea of a normal distribution being standard and the standard normal distribution has a mean of 0 and a variance or a standard deviation of 1. If you have a variance of 1, then the standard deviation is 1, and vice versa, because 1 squared happens to be equal to 1. Then we have the gamma distribution, that scary thing that included the gamma function, and its mean and variance are actually fairly simple. So given the parameters, alpha and beta, the mean of the gamma distribution is alpha divided by beta, and the variance is alpha divided by beta squared, and it's similar with the beta distribution, which has a mean of alpha divided by alpha plus beta, but the variance is slightly more complicated. So given all of these findings, which you can replicate yourself or look them up, we're actually going to conduct an experiment using the R programming language to see whether some of the values generated by, let's say, a normal distribution actually match the theory. So we're going to take a look at whether the calculated values of mean and variance match the kind of predictions that have been made on these slides.

Demo: Mean and Variance

All right, so what we're going to do right now is we're going to actually generate some random numbers. We're going to generate normally distributed random numbers and we're going to specify explicitly the mean and the standard deviation that we want for these numbers, and then we're going to calculate the mean and the variance, and by extension the standard deviation. We're going to calculate this ourselves, and we're going to see whether the calculations actually match. So first of all, let's have a variable, let's call it N, that's going to define how many numbers we want. Let's say 1, 000, like so, and then let's have the random variable X. So that's going to be our norm, and here we specify the number. We specify the mean, so I'm going to have the mean being equal to 3, for example, and the standard deviation is going to be 5. Okay, so the first thing that we do is we calculate the sum of these values. So we can say S is going to be equal to the sum of all the values in X, and then we can calculate their average, so the mean. And I'm actually going to just paste things. So I'm going to say that we expected a mean of 3, and what we got, what do we get? Well, the mean, so I'm going to call it average here, the average here is going to be S divided by N. So the sum of all the numbers divided by the variable N, and that's what I'm going to stick in here, so the average goes in here. Okay, we can already run this and see what we get. Okay, so we expected 3, as you can see here, but we got 2. 72, and as you run it again, you're going to get a different value, obviously, so let's try running it again. We get 2. 92 and 3. 02, so it's close enough, and as you increase the N, you're going to get closer and closer to a 3. 0. All right, so what about the variance? Let's calculate that. So essentially, what we want to do is we want to take the variable X, and we want to generate a new set of values, let's call it Y, and Y is going to be X minus their average squared. So this is what we're getting, and let's once again inspect the values. So we expected variance of 25 because the standard deviation is 5, so the variance is 25, and what we got is the actual expected value of Y. This time around I'm just going to call the function mean on Y. Notice there is a function for everything, including calculating the average of a couple of values. So here let's run this once again, and as you can see, we expected a variance of 25, and we got 25. 32. So as you can see, we're getting roughly the kind of values that you would actually expect.

Joint Distributions

We're now going to discuss the idea of joint distributions. So what are joint distributions all about? Well, sometimes we're interested in not just a single random variable, but several random variables which are related to one another. So for example, let's suppose that you pick a random family out there and you measure their earnings and how many cars they have. Those things

might be related, in fact, I would expect them to be related because, obviously, the more you earn the more likely you are to be able to afford that second or third or whatever car. So these are two random variables which can be in some sort of a relationship. So instead of one variable, we actually consider two or more variables, and we want to have the mathematical apparatus to actually discuss this idea of two or more variables being involved with one another in some calculations or other. Now we're going to focus on two random variables, but the theory that is presented here can be expanded to three or more. So let me show you an example of a joint distribution. So let's suppose we roll 2 dice instead of just a single die, and we want to find the probability that the first die is a 1 and the second is a 3. Now it's a very simple calculation. We already know what it is, so the probability of X being equal to 1 and Y being equal to 3, if we kind of associate X with the first die and Y with the second die, is exactly 136 because we know that these 2 die together generate 36 different values. So a general calculation can be expressed as something called a joint probability mass function.

Probability Mass Function

So a joint probability mass function, if you have two discrete random variables, let's call them X and Y, their joint probability mass function is defined as the probability that the first random variable has a particular value and the second random variable also has a particular value. So the comma between X = x and Y = y actually means and. It means that both of these things have to hold. X has to be the value of small X and Y has to be the value of small Y. Now of course, since we have this idea of total probability always being equal to 1, the joint probability mass function needs to satisfy this requirement that if you sum up all of the joint probabilities around all of the different possible values of X and Y, that you cover the entire sample space, and therefore, the sum probability is equal to 1. Now we need to talk about something called the marginal probability mass function. So essentially, when we get the joint probability mass function, it basically has all the information regarding the distribution of both X and Y, but they're kind of entangled, and we might want to get information about just the distribution of X, and this is where we can try to extract it from the overall probability mass function by fixing the Y values and letting X vary. So the probability with a subscript of X, as you can see here, is actually the marginal probability of X having a particular value, and it's simply the sum across all the different values of Y so you have a sum with a subscript of y index i of all the probabilities where X has a particular value, small x, and Y has that particular value. So you take every single value of Y and you add them up while having the X as the variable. So we call P subscript capital X of X the marginal probability mass function of X. So let's take a look at an example of calculating their

marginal probability mass function with a set of concrete values. So let's suppose that we have random variables X and Y, and they have the joint probability mass function that is shown below, so here you can see the exact probabilities given different values of X and Y respectively, and what we can do is we can try to find the marginal probability mass function for the variable X. Now what do we do? First of all, we calculate the marginal value for the value of 0, so we plug in the value of 0 and we get 1/6 plus 1/4 plus 1/8, which is equal to 13/24. So that is the marginal probability function given that X is equal to 0. Now who wants the same for X equal to 1? We don't even have to do that many calculations because we know that their sum is equal to 1, so we simply say 1 minus the marginal at 0, and that is equal to 11/24 if the first one is 13/24, because remember, they have to add up to a 1. And now that we have this, we have the marginal probability mass function for X, so when X is equal to 0, the value is 13/24. When X is equal to 1, the value is 11/24, and of course, it's 0 in all other cases.

Functions of 2 or More Random Variables

Now we've talked about functions of random variables, but this time around, we're going to talk about functions of two or more random variables, which is a bit more complicated. So sometimes, you have a function which involves two or more random variables. So for example, if you roll two dice and you add up the values that they generate, so you calculate the sum of the values on the two dice, that's a function of two random variables at the same time. So you can describe it as R of capital X and capital Y. Now we have the Law of the Unconscious Statisticians for two or indeed more discrete random variables. So here the expectation of that random variable is the sum where the random function is applied to the values of X and Y multiplied by their joint probability function, and X and Y, so it's kind of like a LOTUS for a single random variable, but here you simply plug in both of the values at the same time. As you may have guessed, we also have this idea of a joint probability density function for continuous random variables, and this, it brings in the idea of variables being jointly continuous. So two random variables X and Y are jointly continuous if, and here comes a really formal definition of the whole thing, if there exists a non-negative function, f subscript XY, such that for any set A in the set of real numbers by real numbers, the probability of X and Y being members of that set is the double integral of the non-negative function that we've mentioned with respect to both X and Y. So the function f with subscript capital X capital Y is called the joint probability density function of X and Y, and just like with the sums, we require that the double integral from minus infinity to infinity of that function, with respect to both X and Y, is equal to 1. So the joint PDF definition might have been a bit too complicated, what with all the abstract set notation and whatnot, so let's take a look at an actual

example, which will hopefully help you understand what's going on here. So let's suppose that X and Y have a joint probability density function of x + c y squared. So this essentially means that instead of just individual probability density functions, the result of the probability density is entwined between the variable x and the variable y. Now let's suppose that we want to find the constant c. Now remember this requirement that the double integral actually has to be equal to 1, this is what we apply here, so we calculate the integral and we start out with minus infinity to infinity of the joint PDF, dx dy, but of course, x and y are, in our example, only defined from 0 to 1. So we do a double integral from 0 to 1 of x + c y squared, and we end up with 1/2 plus 1/3 c being equal to 1, and this easily tells us that c is equal to 3/2.

Marginal PDFs

Just like with the marginal probability mass functions, we have this idea of marginal PDFs, so marginal probability density functions. So the idea here is that instead of taking the double integral, you integrate out the part that you are not interested in. So if you want the marginal PDF for X, you calculate the integral of the joint PDF with respect to Y, and similarly, if you want the marginal for Y, you integrate it with respect to X. And this approach actually lets us find a probability density function for 1 of the 2, or indeed, more variables, given that we have their joint PDF. So let's take a look at an example of calculating a marginal probability density function. So let's suppose we have the joint probability density function where we have f of x, y equal to 10x squared y with the constraint that 0 is less than or equal to y, less than or equal to x, less than or equal to 1. Now this constraint is a really interesting constraint in terms of the integration limits, because remember, we cannot say that y exists between 0 and 1. We have to say that y is between 0 and x, and similarly, x is between y and 1. So let's try to find out the marginal PDF for x and for y as well. So take care of the domains of x and y in this particular case because to find the marginal PDF for x, we integrate from 0 to x of 10x squared y dy, and that gives us 5x to the fourth. So notice that the upper limit of integration is actually x. It's not the value 1 or anything. So as you get the marginal PDF, x itself becomes a member of that PDF even though inside the actual function, inside what's under the integral, it actually kind of disappears as a constant effectively. And the same goes for y, so you integrate from y to 1. Once again, notice that y is the limit of integration here, and once you calculate all the results, you get a function of y. You get 10/3 times y multiplied by 1 - y cubed.

Covariance and Correlation

So one of the problems in statistics is to look for related variables and answering questions whether, in fact, two variables are related. So for example, you want to find out if smoking causes cancer. Not a very positive thought, but we do want to know, so how do you actually prove that smoking does cause cancer? Well, one way of doing it is to conduct measurements, to take measurements of cancer, first of all, so you measure how many people, for example, have cancer as opposed to don't have cancer, and you can be more specific about how advanced their cancer is, and so on, and then you can ask them about the amounts that they actually smoked prior to getting cancer, for example, and that way, you will have two data sets, one data set for whether or not they have cancer and another data set for the amount they smoked. And now what you want is you want to find out whether there is any relationship between those two variables. So in actual fact, there might be other variables that influence the result, but we're going to ignore this idea for now. In real statistical research, you would ask them lots and lots and lots of questions so you can factor out any other variables which can influence that relationship and actually affect it somehow. Anyways, what we want to find out is we want to find out whether the two sets of values actually vary together. So is it the case that if you smoke more, you are more likely to get cancer? So, that's what we want to find out, and that is done using, in very simple terms, the calculation of something called covariance, but of course, one thing about the covariance is it's not very standardized and it kind of scales with the size of the variables themselves, so then we try to get a standardized measure of how related two variables are, and this is something that's called correlation. So in this discussion, we're going to take a look at both covariance and correlation, how they are calculated, and what they're actually used for. All right, so we're going to begin with covariance. So if X and Y are random variables with finite means, which we're going to denote as mu X and mu Y, the covariance of X and Y is defined by the expected value of the product of X minus the mean X and Y minus mean Y if, of course, the expectation exists. So let's take a look at an example of calculating the covariance. Let's suppose that we have the x and y variables and they have a joint probability density function of 2xy + 0. 5, and both x and y are defined in the region from 0 to 1. So to compute the covariance of X and Y, what we have to do is we need to, first of all, calculate the mean, and the mean here is going to be the same whether we calculate it for X or for Y because this mean is for a symmetrical function. So we take the integral from 0 to 1 of both x and y of the probability density function, and we get an actual value. We get 7/12. So this is something that we can use in the calculation of the covariance because then the covariance is just that integral of x - 7/12 multiplied by y - 7/12 multiplied by the original probability density function. Now this is a particularly tricky integral that you probably want to calculate using something like Mathematica, or some symbolic manipulation package, just because it's not the easiest of integrals, shall we say, but the answer is 1/144, so that is the

covariance between variables x and y. Now we have discussed how to calculate covariance, but we haven't discussed what covariance actually is. So covariance is quite simply a numerical measure of the degree to which X and Y vary together. In other words, a covariance is a number which tells us that if X gets higher, does Y also get higher as well? If X goes to low values, does Y also go to low values together? Now the problem with the covariance is that the magnitude of the value calculated in the covariance is actually influenced by the variables themselves, and so what we need is we need something else, and that something else is called correlation. So correlation is kind of like covariance, but it's standardized. It's a finite number, just like probability is, which gives us a better representation of how strongly the variables X and Y vary together. So here is the definition of correlation. Given random variables X and Y with finite variances, sigma squared X and sigma squared Y, the correlation of X and Y is defined as the covariance of X and Y divided by sigma X sigma Y. Now remember, sigma squared is the variance, so sigma is the standard deviation, so we calculate the covariance and we divide it by the product of the standard deviations of the variables X and Y. So let's discuss correlation values. So they are constrained to a range from -1 to +1. Now a value of 0, and notice I'm using the variable rho here, that's what's typically used for the correlation in a population as opposed to a sample, so in a population, the value of 0 means the variables are uncorrelated, they are not related to one another. So, you would expect this value for independent variables. So for example, if you were to generate 100 random values and another 100 random values, the correlation between those 2 data sets should be about 0. It might not be exactly 0, but it should be very close. Now a correlation value of 1 means that as 1 variable grows or shrinks, so does another, and the value of 1 means that it's happening perfectly. A value of smaller than 1 means that there is not such a great correlation, and typically, what we have is we have a whole spectrum, so something above 0. 6 would be called strongly correlated than above, close to 1 would be almost perfectly correlated, and something that is like 0. 2, for example, could be said to be weakly correlated or almost uncorrelated, and so on and so forth. It's worth noting that correlation only measures a linear relationship between the variables, which means that if one variable is affecting another variable in a kind of y equals x squared kind of way, you're not going to be able to capture it using a correlation measure. You're going to have to find something else. Now similarly, if you have a negative correlation value, that means the variables have an inverse relationship. So as one grows, another shrinks, and once again, a -1 means that they do as in perfect lockstep, so as you add 1 unit to 1 variable, you kind of subtract a unit from another, but generally, you're not going to have +1 or -1 as correlation values. You're going to have values which are all the way, all across the spectrum, from -1 to +1 depending on what you're measuring, depending on whether there's any

statistical kind of influence of one over the other, or whether there's any kind of relationship between them.

Demo: Covariance and Correlation

All right, so what we're going to do is we're going to try and stick a couple of values into calculations of covariance and correlation, and we're going to see what you actually get, so first of all, let's start with a bit of a counter example. We're going to generate just two random arrays. So I'm going to have an array with just 1, 000 randomly generated values, normally distributed, but the distribution doesn't really matter here, and let's have another array of pretty much the same. Now as you may have guessed, these arrays are unrelated. They're just random data, so there is no way they have any covariance in them, is there? Well, let's try and find out. So here I'm going to say, paste, so the covariance of X and Y is equal to, and we can actually calculate it using the cov function which is built into R, so just stick the X and Y arrays in here, and let's execute this. So as you can see, the covariance between 2 sets of completely random data is not a complete 0. In fact, it's 0. 00888 whatever, but it's close enough to 0 to say that there is a really no relationship between the 2. Now let's create variables X and Y, which do have a relationship. So let's suppose, for example, that we kind of replicate this, so we have X as a bunch of random values, and then we have Y as, let's say, 5 times X, for example. So now they are in a linear relationship, and as a result, we should get a reasonable covariance value here, so let's stick this in here. So let's take all of this and let's execute this. As you can see the covariance value is now positive. It is 4. 7. Now it doesn't really help us that much because the covariance value here is strongly related to the multiplication factor. So for example, if I put, let's say, an 8 in here, well, you can guess that the covariance is going to be somewhere around 8, in this case, 7. 7. So if we want a standardized value, we need a correlation. So here let's paste the correlation between X and Y, and once again, we have a function called cor, which calculates the correlation. So here X, Y, and that should give us a standardized value somewhere between 0 and 1. In this case, I'm hoping for 1 because they have a perfect linear relationship. So let's execute this, and as you can see, we're getting an exact value of 1 here. Now if I change this from, let's say, a linear relationship to a quadratic relationship like so, let's get rid of the 8 here, then the correlation value is also going to be fairly useless to us. It's not going to give us anything meaningful. Let's actually run this again. As you can see, the correlation value is 0. 03, so it does not show up as an indicator of any relationship except the linear relationship. So we have a quadratic relationship, which is a perfect relationship, but here, it's giving us a value which is close to 0, which is just an indicator of the fact that correlation's only useful for linear relationships and not for other relationship types.

For quadratic and other relationships, we have more sophisticated mathematical apparatus to detect those types.

Summary

All right, so let's try to summarize the things that we've learned as part of this module, and we did learn quite a lot. So we saw that a random variable typically has an expected value or an expectation, which we denote as E of X, which is also the mean or the average. We discussed the idea of variance, which measures how spread out the values are. We looked at functions of two more variables and the fact that they can have a joint probability mass function or probability density function or cumulative distribution function. We discussed the idea of marginal functions, which can be extracted from a joint function to represent a function for one of the particular variables, and we looked at covariance and correlations, which are measures of a linear relationship between two variables.

Looking at Some Special Statistical Distributions

Overview

In this module of the course, we're going to take a look at some special statistical distributions. So, the goal of this course is, once again, fairly simple. We're going to explore some of the more exotic aspects of certain statistical distributions, and here's what we're going to take a look at. So first of all, we'll take a look at the Bernoulli Distribution. We'll discuss the Poisson Distribution. We'll take a look once again at the Normal Distribution to look specifically at one of its aspects. We're also going to discuss the Lognormal Distribution, and we'll finish off the module with a discussion of the Multinomial Distribution.

Bernoulli Distribution

All right, so let's start out by discussing the Bernoulli Distribution. So, let's imagine that you are treating patients. You work in a hospital and you give them some treatment, and this treatment can either succeed or fail, either you see that the treatment is having an effect or it doesn't. So

essentially, we can codify this. We can say that here is a random variable X, and we're going to say it's 0 if the treatment fails and it's 1 if the treatment is successful. So, the probability of the success occurring is being called p here. So, we're going to set the probability value of p, and that's going to indicate whether or not the value is in fact equal to 1. So, p is the probability that X takes on the value 1 using our codify system. And so, what we can do is we can then talk about the collection of all the distributions, with p varying somewhere between 0 or 1 as being the family of the so-called Bernoulli Distributions. So, you can set your own value of p, depending upon the situation. Now, let's suppose, for example, that under this assumption, so we have the probability of X = 1 being equal to p. We have the probability of X being 0 equal to 1- p, obviously. We can now specify a probability function for this variable. So, we can say that f of x is equal to p to the x times 1 - p to the power of -x, and this is only defined for values of X of 0 and 1, obviously, because that's the only values that x can actually take. So, we can calculate some of the descriptive statistics like the expectation, for example. So the expectation for this Bernoulli random variable is equal to p, which is fairly obvious. It's also a fairly intuitive kind of result. The expectation of the value squared is the same as the ordinary expectation because, remember, we have X defined only for 0 and 1, and those 2 numbers, 0 and 1, they have this particular property that if you square them, they don't actually change. So X is in fact equal to X squared, and as a result the expectation of X squared is the same as the expectation of X, which is equal to p, the probability value. Now we can calculate the variance using those two results. So the variance is the expectation of x squared minus the expectation squared. So that's p - p squared, you can factor out the p, and you end up with p multiplied by 1 - p. And we can also talk about the longevity function, which is the expectation for the exponent of t times x, and if you write this out, this becomes p times e to the t + 1 - p.

Bernoulli Trials

All right, now we're going to talk about something called Bernoulli trials. So essentially if you are given a set of random variables, so not just a single random variable, but you have a couple of random variables, X 1, X 2, and all the way up to X n where each of these variables has a Bernoulli distribution, and by the way, notice the notation that I'm using here. So I'm saying X subscript i, and then I'm using a tilde, and I'm saying Bernoulli of p. This is a standard notation that people use to describe that a variable has a particular distribution. So in the case of Bernoulli, you would write out the word Bernoulli, for example, in the case of let's say a normal distribution, we would just put the capital letter N, and then once again in the round brackets you would specify the parameters of the distribution. So here we have a set of random variables from X 1 to X n where

each one of them is distributed with a Bernoulli distribution with a parameter of p, which is the probability value. And the terminology here is that we call this sequence of values, we call them Bernoulli trials with parameter p. So they're kind of like trials in an experiment. Now an infinite sequence of Bernoulli trials is called a Bernoulli process, and in actual fact the term process is also one that is used in statistics quite a bit. Now if you have, let's say you have a coin, and you are tossing this coin, and it's a fair coin. So it has a probability of one-half on landing on either the heads or the tails. So in this case what you're doing is you're generating Bernoulli trials with p equal to one-half. So we've already talked about binomial distribution at some earlier module, but now we're coming back to it, and the reason we're coming back to it is due to an interesting result. So if you take a number of Bernoulli trials, and you calculate their sum, so you take the value of each of the random variables, and you actually have them up together generating a new random variable, then this new random variable has a binomial distribution with parameters, and then p where once again n is the number of trials and p is that probability value. So for a coin toss, for example, it will be one-half, but this really depends on the kind of simulation that you're doing. So for this definition we can easily derive some of the descriptive statistics, some of the measures of this distribution. So for example, we can calculate the expectation. Once again, this is using the result that the expectation of one variable is simply, because it's a linear combination, you can simply specify it as a sum. So it's a sum with i going from one to n of the individual expectations of x i, and obviously we know that the individual expectation is p. So if you take n of them, you end up with n p, and in a similar fashion, we can produce also the variance, which is equal to n p times 1 - p, and the moment generating function, which happens to be the product of the expectations of the exponent of t by x subscript i, and that is equal to p e to the power of t + 1 - p all to the power of n. So let's take a look at an example of a Bernoulli trial. So let's suppose that you have a part which is produced on the conveyor line, and let's say this part has a 10% chance to be defective. So you sample 10 parts from the production line, and the question is how many parts would you expect to be defective? Now I'm sure you've already guessed the answer because it kind of makes sense from an intuitive perspective, but we're actually going to apply the theory that we've learned here, because what you're essentially getting is you're getting a Bernoulli trial with p being equal to 0. 1, and n being equal to 10. So you're getting 10 trials with a probability value of 0. 1. So the expected value here, the expectation of the variable X is equal to 0. 1 times 10, which is equal to 1. In other words, you would expect 1 part in those 10 parts to be defective, which makes perfect sense. If you read the problem statement, it's very intuitive, but here we're doing it in a mathematical way.

Poisson Distribution

Now we're going to talk about the Poisson distribution. Now what we're interested in in the context of the Poisson distribution is measurements of the number of occurrences within a single time period. Now when we talk about occurrences, it can be virtually anything. So for example, it can be the number of customers that arrive to a store, it can be the number of calls to a switchboard, it can be the number of hurricanes or some other extreme weather events, it can be virtually anything, and the Poisson distributions, they model the number of occurrences within a fixed time period. Now the Poisson distribution is interesting because it can also be used to approximate binomial distributions for very small success probabilities, and the reason why it's relevant is because sometimes you do get small success probabilities. For example, if you work at a store and you were trying to calculate the probability of a customer coming to a till in a particular, let's say within 1 minute, then throughout the span of the entire day, that probability can be very, very small, and as a result you have certain mathematical simplifications, and we're even going to take a look at an example where we run those numbers through the R programming language and we see the output. But first of all let's go through the theory. Okay, so let's suppose you are, in fact, a store owner, and you want to model the distribution of the random variable X, which specifies the number of customers arriving in a particular time period. Let's say 1 hour, for example. Now you can model the arrivals in different time periods as being independent. In other words, you are saying that the number of arrivals within 1 hour is independent from the number of arrivals in a different hour. Now let's suppose that, in general, you see five customers an hour on average in your store. So strictly speaking, you can model this almost per second if you wanted to. You could say that, for example, you have an arrival rate of 0. 00138 customers per second, if you decide to actually split it into individual seconds. Now during each second, you either get a customer or you don't get a customer. So you basically have this very simple kind of distribution. You either get a 0 or a 1 with a certain probability of 0. 00138. So this is, obviously if you were to add this up for let's say 1 hour, you would end up with the binomial distribution with n, which is the number of periods you're adding up, being equal to 3, 600, because that's how many seconds you get in 1 hour, and the probability is 0. 00138. That's the probability of somebody coming in within a single second. All right, so you probably want the probability function out of this, but calculation of a probability function with this setup is rather tedious. So this is where the Poisson distribution jumps in, because you get to make certain simplifications. For example, one argument that you could make is that this process does, in fact, have a memory of sorts in that successive values are closely related. So you could say that there is a relationship between the probability of getting a person to come in this second and the

previous second. So if you were to go and do this assumption, you could try to represent the probability function ratios, so f of x + 1 divided by f of x, and you can write those out for the probability distribution because obviously we know the binomial distribution probability function, so we can write them on the top and the bottom, and you end up with n - x times p divided by x + 1 multiplied by 1 - p. Okay, so this is what we have to work with, this is the thing that we want to simplify, ideally. So how can we simplify this, you might ask? Well, remember the whole point here is that some of these values, and certainly the probability, is actually very small, so we can make two simplifications. First of all we can say that for the first few values, let's say for the first 30 values in this particular case, the value of n - x can be taken to be approximately equal to n, simply because x, for the first couple of values, is going to be rather small. It's going to be tiny, and the second is that dividing by 1 - p actually has very little effect, because remember, p, the probability of somebody coming in in a given second, is actually a tiny value. So 1 - p is almost the same as 1, so you're basically kind of canceling out of that particular term there. So we can approximately rewrite 10 - x times p divided by x + 1 times 1 - p as simply n p divided by x + 1. So that's a certain improvement, shall we say? Now we can introduce a new term, lambda, which is going to be equal to n times p, and we have the following interesting recurrence relationship. So we have the f of x + 1 being approximately equal to f of x times lambda divided by x + 1. Okay, so what can we do with this recurrence relationship? Well we can try to expand it out for the different values of x. So for example, if you want to express f of 1, we can express it through f of 0 times lambda. If we want f of 2, we can express it through f of 1, but remember, f of 1 itself is expressible through f of 0, and as you continue to expand these different values, you end up with something interesting on the right-hand side. So you end up with the f 0 term on the right-hand side, but you're also accumulating the lambdas in the enumerator and some values in the denominator, and then actual fact we can generalize. We can say that f of x is generally equal to f of 0 multiplied by lambda to the power of x, whatever the value of x is, divided by x factorial. Notice how you have half, third, and so on being multiplied all together, so you have division by 1 times 2 times 3 times 4, and so on, so that's the factorial, basically. So we want this to become a valid probability function. We actually want a valid probability function that has a sum which is equal to 1. So we want this whole thing, f of 0 multiplied by lambda to the x divided by x factorial, we'll want it to be 1 if we sum it across all the different values of x. So x equals 0, 1, 2, and all the way up to infinity. So how can we get this? Well, to guarantee that the sum is 1, we can set f of 0 to 1 divided by this whole sum, 1 divided by lambda to the x divided by x factorial. Now interestingly enough, that sum at the bottom actually happens to be the expansion, the series expansion, for the exponent. So effectively the sum is actually equal to 1 divided by e to the lambda, or in other words e to the minus lambda. So we found our f of 0, and we can plug it into

the definition of the probability function, and we have the following. So here is the formal definition for the Poisson distribution. So a random variable x has a Poisson distribution with a mean lambda greater than 0 if, and only if, it has the probability function e to the minus lambda multiplied by lambda to the x divided by x factorial for x being equal to 0, 1, 2, and every natural number, including 0. So that's the definition of the Poisson distribution. All right, now here are some descriptives for the Poisson distribution. So if x is distributed with a Poisson distribution with parameter lambda, then the mean is equal to lambda. In fact, we've sort of mentioned the fact that it's a mean on the previous slide, and also the variance is equal to the lambda as well, and the moment generating function is e to the lambda times e to the t - 1.

Demo: Poisson Distribution

All right, so now we're going to take a look at the binomial distribution and the Poisson distribution. We're going to compare them using that example that we had when we were discussing the Poisson distribution on the slides. So here we have 3, 600 different subdivisions. So 3, 600 seconds in 1 hour, and the lambda value in this case is going to be equal to 5, that's the value of the number of customers that arrive at a store during one particular hour. So let's set up the result matrix that's going to be a matrix where the number of rows is going to be 10, so I'll take the first 10 values, and the number of columns is going to be 2, because I want to store both a binomial, as well as the Poisson distribution values. So let's have x in 0 all the way up to 9, and let's actually assign the different parts of the result. So result at x + 1, 1, so we're going to do the binomial first of all. So that's going to be dbinom, and that's how you calculate the binomial function. So here we stick in x, followed by n, followed by lambda divided by n. So that's the first part, and then we're going to do the Poisson as well. So result at x + 1, 2, so that's the second column, is going to be dpois, that's how you call the Poisson function, with x and lambda. And having acquired these values, we can actually print them. So let's print the result, and let's actually run all of this and see what we get. So if I just expand this for a moment, you can see that the results are pretty much identical. You get 0. 0067, 67, 33, 33, and as you follow this, you can see that for the first 10 values, you have very similar results. Obviously they're not identical, but they are very close, they're close up to maybe four decimal places, which is good enough for us.

Normal Distribution Revisited

We've already talked about normal distribution in one of the previous modules, and it's a very important distribution. It's one of those distributions that shows up all over the place, and we

want to revisit it now to mention a few other things related to this distribution. So let's just remind ourselves of the probability density function for the normal distribution. Notice the notation here. So I'm saying x tilde, and then n of mu and sigma squared. So this basically says that x is normally distributed with mean mu and variant sigma squared if, and only if the variable has the following probability density function. So it's 1 divided by sigma square root of 2 pi times the exponent of x - mu squared divided by 2 sigma squared. Sometimes the x - mu over sigma is grouped together and then squared, and the minus one-half gets taken out. So don't be surprised by the slight differences in the way the people write this down. So the expected value of random variable x is obviously mu, the mean, and the variance is sigma squared. The moment generating function for the normally distributed random variable is the exponent of mu t + sigma squared t squared divided by 2, and let's actually go and explore one of the aspects of the normal distribution with this idea of linear transformations. So essentially, if you have a random variable x which has a normal distribution with mean mu and variance sigma squared, then you can define a new variable, let's call it y, which is a linear combination of x and some other constant. So if y is equal to x + b, and a and b are just some constants such that a is not equal to 0, then what we can say is that y has a normal distribution with a slightly different means and variance. So in this case, the mean would be a mu + b where mu is the old mean, of course, and the variance will be a squared sigma squared where once again sigma squared is the variance that we had for the original variable. So for example, let's suppose that you have the variable x normally distributed with mean 0 and variance 1, and standard deviation 1, of course, then if we let y equals to 2 x + 3, then we can say that y is normally distributed with a mean of 3 and a variance of 4, because if you plug in those values, that's what you're going to get. Now let's talk about something called the standard normal distribution. So the standard normal distribution is essentially a specialization of the normal distribution, with a mean of 0 and a variance value of 1, and of course the standard deviation is equal to 1 as well. So this has the probability density function, which is simplified compared to the original normal distribution because some of the terms have been replaced by 0s and 1s respectively. So you have 1 over the square root of 2 pi multiplied by the exponent of -x squared over 2, and similarly the cdf for this is the integral from minus infinity to x of that pdf. Now as you may have guessed, the cdf does not have an analytic solution. There is no analytic integral for e to the minus x squared, and as a result you will have to use numerical methods or use tables to perform the actual calculations. Also note that this function is actually symmetric, and because it's symmetric, the cdf of a negative value, so the cdf of -x is the same as 1 minus the cdf of the positive value, 1 - phi of x, and we're using the letter phi here, the small phi represents the pdf, and the large phi represents the cdf respectively. So something we need to talk about as well is the conversion of random normal distribution. So normal distribution with particular

parameters to the standard normal distribution. So if x is normally distributed with mean mu and variant sigma squared, and if some function F, capital F, is the cdf of the random variable x, then what we can say is that we can introduce a new variable, let's call it Z, and if we take Z to be X minus mu divided by sigma, this has a standard normal distribution. So what this means is that we can transform some arbitrary and normally distributed variable to a standard normally distributed variable, and then it makes it easier for us to perform the calculations, and it's particularly easy if you are, for example, using tables as opposed to the R language, or some other statistical environment, because in the modern caste systems it doesn't really matter one way or another. You can simply just plug in any values and get your result, but this makes it a bit simpler. So essentially, we can specify the new cdf for the random variable x, we can specify it as the cdf for x - mu divided by sigma. So let's take a look at an example where you would convert a normal distribution to a standard normal distribution in order to calculate something. So suppose you have the random variable x, which has a mu of 5 and sigma of 2. So you want to calculate the probability that x lies between 1 and 8. So what you do is you define any random variable, let's go with Z, which is defined as X - 5 divided by 2. So you subtract the mean and you divide it by the standard deviation, and then what you get is you get this new random variable, Z, which is actually a standard normal random variable. Okay, so having set this up, what you can say is that the probability of x being between 1 and 8 is the same as the probability of Z being between -2 and 1. 5. And at this point what you can do is you can perform a few manipulations in order to actually make this value even easier to calculate, and these manipulations are obviously not required if using a computer, but if you're using tables, and lots of statistic books comes with tables nowadays, you have to know how to perform these kinds of manipulations. So essentially we're saying that in order for the random variable Z to be somewhere between -2 and 1. 5, that is effectively the probability that the value is less than 1. 5 minus the probability that the value is somewhere to the left of -2, so you subtract that from the other, and of course here we can plug in the function phi, which is the cumulative distribution function. So we calculate the cdf of 1. 5 minus the cdf of -2, and of course we have this symmetry relationship. So phi of -2 is the same as 1 - phi of 2, so you get phi of 1. 1 minus brackets 1 - phi of 2, and these are the values that you can simply look up and you get the final value. So the probability of x being between 1 and 8 is somewhere around 0. 91.

Lognormal Distribution

All right, so one thing that I want to talk about with a connection to the normal distribution is something called the lognormal distribution. So let's suppose that you have random variables X

and Y, such that Y is equal to the logarithm of X. Now if we find that Y has a normal distribution with mean mu and variance sigma squared, then X is said to have something called a lognormal distribution. So from the definition of the normal distribution, we can get the moment generating function of Y as the following expression. So we get the psi of t as the exponent of mu t + one-half sigma squared t squared. However, there is a bit of a simplification here that we can apply to make it easier for us to calculate things like the expected value and the variance of the lognormal distribution. So if you look at the lognormal distribution carefully, you'll see that it contains a log. I mean, it's fairly obvious. So the moment generating function, of course, involves the exponent. So when you put the exponent and the log together, they are supposed to vanish, right? And that's exactly what happens. So the psi of t, the moment generating function for the random variable Y, is the expectation of e to the power of t times Y, but of course Y itself is log X, so you have e to the power of t times log X, which becomes the expectation of X to the power of t. Now this is a very interesting result. Basically what it says is that if you want the expectation of X to the power of something, you can calculate the moment generating function with that particular value. You can simply take the value of t, and you can stick it into the expression that we have just above, so the moment generating function written out explicitly. And as a result of that, we can easily calculate, for example, the expected value. So the expectation of X is really the expectation of X to the power of 1, and according to the result above, this is equal to psi of 1. So you take the value of 1 and you stick it into the expression above where you have the t and the t squared, 1 squared is obviously 1, so you end up with the exponent of mu + one-half sigma squared. And the same goes for the variance, because remember, the variance is the expectation of x squared minus the ordinary expectation squared. So that's psi of 2 - psi of 1, you plug in those values and you get this expression. So you have the exponent of 2 mu + sigma squared multiplied by the exponent of sigma squared - 1.

Multinomial Distribution

The last thing that I wanted to mention in this module is the multinomial distribution, and as you may have guessed, this is a generalization of the binomial distribution. So in the binomial distribution, for example, you have just two possible outcomes, it can be yes or no, it can be black or white, or something like that, but sometimes you have more than two outcomes. So for example, let's suppose you're measuring blood types in people. There are four different blood types, so essentially you need a different model, you need to expand effectively this model to support the idea of what is the probability of let's say sampling a group of people, and having let's say half of the people with type O and a quarter of the people with type A, and so on and so

forth. So you need this expansion to the model, and this is what multinomial coefficients actually give us. So given that you have a couple of items, x subscript i, which has a certain type of criteria. So let's suppose that you have k different criteria, then the multinomial coefficient, which is also, by the way, written in brackets, just like the binomial, is simply the factorial of the top divided by the product of the factorials of the other elements. And by the way, even though I am having n over x 1, x 2, x k, we don't often put in the commas, especially when it comes to a small number of coefficients. Instead of that, we actually put spaces in between or somehow try to spread them out. The commas aren't really used, but they're used here for illustration because I want to illustrate the fact that there is a sequence, x 1, x 2, all the way up to x k. So if you have a random variable, x, which actually is not a variable if you have a vector of the different random variables representing the counts, let's say the counts of people with different blood types, and you have the small x representing the possible value, possible random vector from a selection of all the possible random vectors, then the join probability mass function of that vector is the multinomial coefficient, so the thing with n on top and x 1 all the way up to x k at the bottom multiplied by the individual probabilities. So in the case of blood types, for example, if you knew the probability of getting the different blood types, they would be the ones that are multiplied here, and notice every single one of them is taken to the power of x, is taken to the power of that particular value. So let's take a look at an example of a multinomial distribution. So let's suppose that you are choosing a couple people who all watch television. Let's suppose that 23% of people watch television for 0 to 10 minutes, and 59% watch between 10 minutes and 1 hour, and 18% watch for more than an hour. Now you can ask questions such as given a sample of 20 people, what's the probability that 7 people watch 10 minutes or less of tv, and 8 people watch between 10 minutes and 1 hour, and the remaining 5 obviously have to watch more than an hour? What the probability of all of those things occurring? And this is where you can simply plug in the values into the multinomial probability density function, that joint function that we had in the previous slide, and you get 20 factorial divided by 7 factorial times 8 factorial times 5 factorial, because that's how many people you actually have if you take the first 15, and there's a total of 20, you end up with 5, and then you have each of the probabilities taken to the power of how many people you actually expect there to be, and you get 0. 00942, which is a rather small probability value, and it also makes sense because there are so many probabilities when you are sampling 20 people and you are asking them how much tv they watch.

Summary

All right, so let's try to summarize some of the things that we've learned in this module. So first of all, we looked at the Bernoulli distribution. We saw that it's a distribution which allows you one of two different possible values with a sudden probability p. We saw that a sequence of those values are called Bernoulli trials, and we saw that the sum of Bernoulli trials actually gives you a binomial distributed random variable. Then we looked at the Poisson distribution, which serves as a good approximation for the binomial distribution for small probability values. We looked at the normal distribution once again, and we looked specifically at the linear transformation, as well as the lognormal distribution, and finally we talked about the multinomial distribution, which is just like the binomial distribution, except that a variable can take on more than two different values.

Course Summary

All right, now that we've reached the end of the course, let's actually talk about some of things that we've learned as part of this course. So the first module of this course introduced this idea of probability, and in simple terms, probability is quite simply a number between 0 and 1, which represents the likelihood of something happening. We discussed some of the terminology that's used in statistical research, such as the idea of experiments and events. We had a brief primer in set theory so that you could reason about events overlapping, for example. We talked about the idea of sample spaces. And then we looked at applications of combinatorics for actually counting the number of elements in a particular sample space. The next module discussed the idea of conditional probability, that is the probability of events provided that some other events have occurred, and we discussed the idea of events being dependent or independent, we looked at the Bayes Theorem, and its application to a well-known problem known as Gambler's Ruin. In the next module, we talked about the idea of a random variable, and the idea that a random variable has a distribution. We discussed the difference between discrete random variables, which can take a finite number of different states, as opposed to continuous random variables, and we looked at their descriptive statistics, such as the probability mass function, the cumulative density function, the probability density function. We covered some of the key distributions, which are often used in statistical research, such as the normal distribution, for example. In the subsequent module, we discussed the idea of expectation, and we applied the notion of expectation to calculate some parameters and some aspects of a particular distribution. We talked about the mean, the variance, as well as different moments of a distribution, and we talked about the idea of a moment generating function. So in the final module, we talked about special distributions, a couple of distributions which are not as frequently used, but also important to know about. So we looked at those, and this concludes this particular course.

Course author

Dmitri Nesteruk

Dmitri Nesteruk is a quantitative analyst, developer, speaker, and podcaster. His interests lie in software development and integration practices in the areas of computation, quantitative finance,...

Course info

Level	Beginner
Rating	★★★★★ (49)
My rating	★★★★★
Duration	4h 24m
Released	28 Mar 2018

Share course