**IOC Topic 11b – Advanced Data Science**

Transcript & Notes: PART 4

Author: Dr. Robert Lyon

Contact: robert.lyon@edgehill.ac.uk ([www.scienceguyrob.com](http://www.scienceguyrob.com/))

Institution: Edge Hill University

Version: 1.0

**Topic 11b, Part 4**

**Introduction Slide**

Hello and welcome to Part 4 of Topic 11b, Advanced Data Science. During this topic I'll introduce what data science is, the basic principles underpinning data science, and some important data science tools that may be unfamiliar to you. My name is Dr. Robert Lyon, and I’ll be taking you through the learning material.

**Slide 1**

What material will we cover while studying this topic? Well, this topic aims to introduce…

* What data science is all about.
* Key concepts underpinning good data science – primarily the scientific method.
* Useful terminology that will help you navigate the world of data science.
* Important tools crucial for successful and reproducible data science – these are the tools provided by Statistics.
* Data collection & Experiment Design practices.
* Probability basics – very important for statistical inference.
* Data distributions that describe the characteristics of data.
* Hypothesis testing – a formal method for testing predictions.

The aim: to help you understand what it means to be a data scientist and to get you familiar with data science tools.

In part 4 we’ll look at probability and data distributions. In data science, we often need to make decisions or predictions using available data. This requires determining which decisions are more likely to lead to the desired outcome, or which predictions will probably hold true in the long run. To do this we need to understand the basics of probability.

**Slide 2**

* Studying probability helps us understand the randomness in the data we collect.
* We often think about randomness in terms of random variables being produced by some random process. For example, will or won’t a customer make a purchase today? Even if a process isn’t strictly random, it can still help to think about it in this way.
* To understand probability we first think a little about it’s nature.
* Just because something is probable, does not mean it will happen. Just as something being rare doesn’t mean it won’t happen.
* Take for instance a simple 6-sided die. If we roll the die then the probability of landing a 6 is (one sixth). So, what happens if we run an experiment to measure the actual outcome? We start rolling a fair die using an unbiased rolling process. After 10 rolls, we record the fraction that produced a 6. We keep rolling and recording the results. Initially the fraction of sixes fluctuates a great deal. As we run more trials, the fraction begins to settle down. Eventually it converges on a value very close to 0.167. This is one-sixth rounded to two decimal places. So we’ve found that in the long run, the probability is very close to what we’d expect – thanks to the law of large numbers. If we kept going, eventually we’d get to an almost exact answer. What’s important to understand is that randomness comes into play, which introduces variability into our results.

**Slide 3**

Before we go any further, please watch the following video (https://youtu.be/uzkc-qNVoOk). It will provide a basic introduction to probability that will help you understand the content that follows.

**Slide 4**

* During our first die experiment, we considered a specific type of probabilistic event.
* Each trial outcome was an independent event. Independent events do not influence the outcome of any other events. So each trial with the die does not affect the next.
* Another way to say this, is that each event is mutually exclusive, which means disjoint – they cannot happen at the same time.
* For instance, I cannot get an odd and an even number during a dice roll – there is only one die and therefore disjoint outcomes. That’s what we can see depicted here. The are two disjoint outcomes – either an odd number will be rolled, or an even number will be. Each outcome is equally likely. We can see how these outcomes are separate and therefore independent.
* The outcome “roll a six and get an even number” are not disjoint events – these can both happen with a roll of the dice. We can visualise this scenario to make things clearer. The outcomes intersect. Thus, the chance of getting an even number and of it being 6, depends on two outcomes. This is why these are dependent events.

**Slide 5**

* We usually express probabilities using notation. We do this as it be difficult to explain in English exactly what we mean when discussing or describing probability.
* At first this notation can appear confusing. I promise it is relatively straightforward once you get used to it. Just keep in mind that the notation is representing numbers that you can do basic math with. With that said, let’s consider some notation.
* An upper case can be read as “the Probability”.
* Usually this probability refers to some outcome or event. To show this, you may use an index such as , which then allows us to neatly describe the probability of eventoccurring via . So the outcome of rolling a 6 using a dice could be described as follows: .
* We can see that probabilities can be described as fractions, or as decimal numbers.
* Sometimes we may see probabilities written like this: . Here ”Event” is a placeholder for any event we can think off. For example, we can write the probability of rolling a 6 as shown: .
* I’ll use this notation from now on, as it’s easier for teaching.

**Slide 6**

* We may also use variables to represent probabilities or events, simply to make the notation easier to interpret.
* For example, we could use the letter A to represent “Roll an odd number”.
* We can see that using variables makes things more concise.
* There are some operators you should know about too.
* First there is the “not” operator represented by the symbol shown: .
* Not is used to negate the probability of an event happening. So, if we define as the probability of rolling an odd number, and the probability of rolling an even number, we have 2 independent events. The probability of happening is 50%. Not describes the probability of everything but happening – in this case that would be the probability of rolling an even number. If doesn’t happen, then it follows that must happen in this case.
* The sum of probabilities for all events should always add up to 1. So, the probability of odd is 0.5, and the probability of even is 0.5, which summed give us 1.

**Slide 7**

* For dependent events, we can describe the probability of all outcomes occurring, or just one of many occurring.
* We can do this using notation for logical AND (). & logical OR ().
* For example, we can define the probability of rolling a 6 and an even number, or the probability of rolling a number divisible by 2 or 3.
* Notice in the example for logical OR, some of the outcomes don’t sit inside either outcomes C or D. These are represented by the notation . If you think about this for a second, it makes sense. Both 1 and 5 are not divisible by 2, and not divisible by 3. Take some time to digest this notation if it doesn’t make sense.

**Slide 8**

With just the simple notation you’ve learned, you can represent lots of different types of probabilistic event. Let’s recap these now. Don’t worry about trying to remember these. Instead I want you to be able to interpret the notation and the diagrams, so that you can model any problems you may face in the real-world. So, we are able to represent.

* The probability of a single event, A: .
* The probability of a single event A not occurring: .
* The probability of A and another event B occurring: .
* The probability of A or another event B occurring: .
* The probability of A occurring and another event B not occurring: .
* The probability of event A or event B not occurring: .
* The probability of event A and event B not occurring: .
* Then there are two more outcomes for you to think about in your own time.

When faced with a probability problem, it can help to visualise it as shown here. It can help you understand what’s really going on.

**Slide 9**

So far we’ve only considered simple events, and no more than two events occurring at the same time. In the real world, even simple problems can become complex to visualise and understand. Consider the following problem. We have a bag of 8 marbles of which 3 are blue, 2 are yellow, and 3 are white. We can define some simple notation to describe the probability of each type of marble being picked. If we draw from the bag, there are three potential outcomes – pick yellow, pick a white, and pick a blue. For each of these outcomes we can draw again. Let’s stick with just one path where we first pick a yellow marble, followed by a white marble. We draw again and get a white marble, and draw one more time and again, get a white marble. How likely was this series of draws? Well, we can use some basic probability theory to work this out. If we multiply the likelihood of each event occurring, then we get a probability estimate. It seems that the sequence yellow, white, white, white, is quite unlikely. There’s only a 0.7% chance of it occurring. You probably have questions – i.e. why are the probabilities multiplied to give the answer. Don’t worry about that for now. Instead focus on the structure of the diagram we created. It has a tree-like structure, thus this is known as a tree diagram. They are very useful for studying conditional probability. This is probability that depends on the outcome of previous events.

You might feel that what I’ve described so far is quite distant and unrelated to real-world data science. Well, imagine swapping the marbles for customers, and imagine the tree diagram now describes user interaction with a website – i.e. Which pages they navigate to and which they don’t. Using these simple tools we can model user behaviour.

**Slide 10**

You may be wondering why we multiplied the probabilities together in the last example. We do this as somebody much smarter than I, figured out the rules of probability which you can prove via experimentation – like we did at the start with the dice rolling trial. To help you understand probability better, we describe these rules.

We have the addition rule for disjoint outcomes, that allows us to compute the probability of 1 or more independent events happening. Here we can see this represented. It shows that the chance of rolling an odd, or an even number with a die, is 1.

The multiplication rule, which we used in the previous slide, allows us to compute the probability of all events happening at the same time. Here we can see an example of how this works. If we have two independent die rolls, we can work out the chances of an even number on the first roll, and an odd number on the second roll. There is a 25% chance of this happening. Is this correct? We can list all the potential outcomes and figure this out for ourselves. We can see that there are 4 possible outcomes, each equally likely with a 25% chance. So, the formula gave us the correct answer.

**Slide 11**

There are some further rules. The addition rule for any outcome allows us to determine the likelihood of at least one of two events happening. For example, what is the chance of rolling a even die, or an odd die? This is exactly 1, as at least one of these will happen.

How about another example. What is the chance of rolling an even die or a 6. Well, the chance of rolling an even die is 50%, and the chance of rolling a six is approximately 16%. If we apply the rule, we find the probability of even die or a 6 is 50%.

Then there is the complement rule. This simple rule tells us that the chance of an event happening or not happening, sums to 1.0. For example, the probability of rolling an even die plus the probability of rolling an odd die, is exactly 1. As both are 50% likely to happen.

**Slide 12**

Let’s reinforce what we’ve learned. Please watch the videos on the slide. First watch the video (https://youtu.be/QE2uR6Z-NcU) on the left-hand size. It will introduce the addition rule again but with further examples. Whilst the second video on the right (https://youtu.be/94AmzeR9n2w), will recap the multiplication and addition rules for dependent and independent events.

**Slide 13**

Conditional Probability is the second to last probability concept we’ll look it. It allows us to estimate the probability of an event occurring, given that some other event has already occurred. I’ve already sneakily introduced this idea using the marble example on slide 8. For instance, what is the probability of picking a blue marble, if a white marble has already been picked?

The probability of a white marble being picked is: . Then we compute the probability of picking a blue marble, which is: . This is the conditional probability of picking blue, given that white has already been picked. Now this is easy to compute in this example. But in the real-world, our problems are much more complex with large tree-diagrams. So rather than creating a tree-diagram, we instead use the formula for conditional probability to work things out. Interpret this formula to mean, the probability of happening, given that already happened. Read the “pipe” symbol (|) that separates and to mean, given .

Now we can use this formula to compute the probability of picking a blue marble, given that we already picked a white marble, as shown. Now this looks like a lot of math for this simple problem – and yes, for small problems we could just use the tree-diagram. But the formula becomes very powerful for complex problems. Finally, we know from an earlier example that we can use the multiplication rule to determine the probability of the events, white and then blue () happening in sequence. It is approximately 21.4%. Remember this is not the same as the conditional probability. Before moving on, take some time to read through the formula and the example on the slide. Try to understand what's happening. Perhaps try to compute the conditional probability of blue given yellow () using the formula to be sure you understand it (answer is 0.5 = 50% ).

**Slide 14**

So far, we’ve considered basic probability. You probably feel 1,000 miles from data science! But you’ve just acquired the foundational knowledge needed to help you answer questions about data, using the data before us as evidence. For instance – before starting this course, suppose you faced question such as – which students are most likely to struggle at A-level based on GCSE results, or which customers are unlikely to pay back their loans? You may have been able to plot some data, compute some statistics, and maybe use anecdotal evidence to make predictions – i.e. poor people are likely to struggle paying credit back. None of this is very exact. But now with probability theory behind you, you can rigorously analyse data and make principled predictions. Let’s try an example.

* Suppose we are government data scientists, put in charge of guiding policy based on data.
* We’re given data describing drug use by students and their parents. There are 4 potential outcomes.
  + Outcome 1 – the student uses drugs & the parent used in the past.
  + Outcome 2 – the student uses drugs & the parent has never used.
  + Outcome 3 – the student doesn’t use drugs and the parent used in the past.
  + Outcome 4 – neither the student or parent has ever used.
* We form a contingency table using the data.
* It shows the number of parents who used drugs in the past, and students actively using drugs now.
* We can visualize the data in the table as shown. We can see that 0.28 = 28% of students use and have parents that used in the past.
* Take some time to study the contingency table – be sure you understand it’s content.

**Slide 15**

* A student who doesn’t use drugs is chosen at random. What is the chance that at least one of her parents used drug in the past: ?
* We can use the conditional probability rule to solve this, or we can look at the contingency table. Let’s set and . We can see that students who don’t use drugs, have parents that have used in the past. So the probability is 37.6%.
* But we can compute lots of other probabilities now too. What about,
  + The probability that the Student uses, and the parent doesn’t use: When ready, pause the slides, and take a moment to tackle this. The answer is 21% - does that match what you got?
  + The probability that the student uses: ? When ready, pause the slides, and take a moment to tackle this. The answer is 49.2%.
  + Finally, the probability that the student uses, given that the parent used: ? When ready, pause the slides, and take a moment to tackle this. The answer is 60%.

**Slide 16**

I summarise the rules we’ve encountered so far here. As you can see, there are lot’s of questions we can now answer using probability. You can now form your own questions, express them using notation, and apply the rules.

**Slide 17**

Conditional probability is very important. Before we continue, please watch the two videos shown on the slide. The left-hand video (https://youtu.be/bgCMjHzXTXs) introduces conditional probability from an alternative perspective. While the right-hand video (https://youtu.be/ES9HFNDu4Bs) will provide some examples of how to use conditional probability to solve problems.

**Slide 18**

We skipped over some notation in the last few slides. We can fill that gap now.

In our last example, we tackled a problem related to estimating drug use. We encountered two types of probabilities that have special names.

* Marginal probability – the probability of just one outcome occurring (over exactly one variable). We can see here that there are three marginal probabilities in the data.
* Joint probability – the probability of two outcomes occurring (over two or more variables). We can see that there is only one joint probability under consideration for this problem.

**Slide 19**

* Sometimes we have data describing a specific situation that doesn’t quite suit our purposes.
* For example, suppose we have a collection of patients being studied during a medical trial with respect to a specific disease. We know how likely it is for these patients to be afflicted by the disease.
* We see that the disease is very rare – there is an approximately 99.7% chance that an arbitrary patient doesn’t have the disease.
* Now suppose I tell you that the aim of the trial is to determine how effective a procedure is for detecting the disease.
  + For those patients with the illness there is an 89% chance they test positive.
  + For those patients with the disease there is an 11% chance they test negative.
  + Whilst for those patients that don’t have the disease, there is a 7% chance they will test positive for the disease.
  + For those patients without the disease there is an 93% chance they will test negative for the disease.
* Given this information, we can calculate the likelihood of a person with or without the disease, testing positive or negative. For instance:
  + The probability of an ill patient testing positive = 0.00312 which is 0.312%.
  + The probability of an ill patient testing negative = 0.00038 which is 0.038%.
  + The probability of a healthy patient testing positive = 0.06976 which is 6.976%.
  + The probability of a healthy patient testing negative = 0.92675 which is 92.675%.

**Slide 20**

* Sounds ok so far – but what if I want to ask a different question. I want to ask – what is the probability that a patient testing positive is ill?
* We can invert the probability tree to ask this question. All the data is there, except it is very difficult to do – it gets complicated.
* The probability of a patient testing positive being ill, is actually just 4% - so how did I compute it?
* I used our final rule from probability theory to compute the likelihood – this rule is called Bayes Theorem.

**Slide 21**

* This simple theorem lets us compute conditional probabilities in an easy way.
* It’s very powerful – it actually forms the basis for a number of machine learning based systems.
* What does it say?

The probability of 𝐴 given that 𝐵 has happened, is equal to the probability of 𝐵 given that 𝐴 has happened, multiplied by the probability of 𝐴, all divided by the probability of 𝐵.

* It is unlikely that this will make sense just by looking at it.
* We can expand it, so it makes more sense, but it’s still a lot to digest.
* So, let’s use it to answer the question we posed on the last slide.

**Slide 22**

I previously claimed that the probability of a patient testing positive and being ill just 4%. Let’s show how I know that. Here we have the probabilities we’ll need. Next, we have Bayes Theorem. Let's take the first term and try to compute this. We can see that this is equal to the probability of and , , divided by the probability of . If we take a look at this for a second, and look back at Bayes Theorem, we can see the common term appears in one as a dividing term, and in another as a multiplicative term. This means we can remove this term to simplify our calculations as shown. Is this the correct thing to do? Well let’s test this will some example data to illustrate the point. If the probability of and is 0.5 and the probability of is 0.8, then the probability of given is 0.625. When we multiply this by the probability of , , back in the Bayes formula, we get 0.5. In other words, it all just cancelled the probability of out leaving the probability of and : .

**Slide 23**

Now that we’ve simplified our equation by cancelling out the some of the terms, we can start plugging in values and compute the numerator in Bayes Theorem. The probability of a positive test result for a sick patient is 0.89. The probability of a patient being ill is 0.0035. When we multiply these values, we can see that the probability of testing positive and being ill, equals 0.00312 which is 0.312%.

Next, we look at the probability of a test being positive (the denominator of the equation). A test can be positive in two different scenarios – when the patient is ill and when the patient is fine. We can further decompose this probability to explain things more clearly. Spend some time looking at this formula. It will seem confusing at first. But then look at the diagram on the left - we can see what the formula is doing. We first take the probability of , and then multiply this by the conditional probability of given We then repeat this but for the half of the equation following the plus sign. We plug in the numbers and compute the probabilities as shown. We see that the chance of obtaining a positive test result irrespective of anything else, is 7.288%.

**Slide 24**

Finally, we can sub in the values we computed previously. This gives us the following answer. Thus the chance of a patient being ill given a positive test result is 4.28%. So you can see, my earlier claim has been shown correct using Bayes Theorem. I appreciate this theorem may still seem mysterious or complicated. Later you’ll get to try this for yourself during some topic activities. This should help make things clearer.

**Slide 25**

When ready watch the video (https://youtu.be/OqmJhPQYRc8) shown on the slide. It will review Bayes theorem again, and provide some further examples of how it can be used. It is important to understand that this simple theorem is extremely powerful. This is why I’ve spent so much time covering the background of probability theory. Even if you don’t understand the theory in detail, you can still use it by plugging values in. That kind of practice will help you learn too – it should help build an intuitive understanding over time.

**Slide 26**

So far we’ve considered probability theory. The next leap in understanding comes when you start to think about probability distributions, or data distributions in general. For example, if we have two dice, when we roll them, the sum of the numbers which come up give rise to a distribution. We can represent this distribution first using a table. We can see how frequent different outcomes are. A sum of seven is the most common outcome – it can be formed via multiple outcomes as we can see. So this is the most likely outcome when rolling two dice. We can also visualise this data using a histogram. Now we can see more clearly, that some summed values crop up more often than others. There is a bell shape to the data. The data is said to be normally distributed. We’ll come back to this very important data distribution next time. For now, it’s important to understand that a simple random process gave rise to a normal distribution naturally. Finally, note that in this case we have a discrete distribution which can be represented via a histogram as shown. This is because the roll of two dice has a fixed and discrete number of potential outcomes. Not all data distributions are discrete.

**Slide 27**

Here we have the same type of normal distribution as before, however we no longer have a histogram. Instead, we have a smooth curve. This represents a data distribution that has an unlimited number of potential values. Each value in the distribution is a real-valued number that occurs with some probability. This distribution could represent, for example, human height. We won’t deal with many continuous distributions in this course. To study them we need to be familiar with more complicated mathematics. Nonetheless I want you to know they exist, so you can be ready for them when they crop up.

**Slide 28**

To help you cement an understanding of what I mean by data distributions, please watch the video (https://youtu.be/bPFNxD3Yg6U) shown on the slide. This will refresh your memory of unimodal, bi-modal, and multi-modal data. It will also introduce you to uniform distributions.

**Slide 29**

Before you go, I have one more video (https://youtu.be/9TDjifpGj-k) to share. This may help make Bayes Theorem more relatable. When ready, please watch the video on the slide.

**Slide 30**

We’ve reached another checkpoint. Let’s recap what we’ve introduced so far.

* The nature of probability.
* Different types of probabilistic event.
* Probability notation.
* How to define events and express them happening independently or together.
* Tree-diagrams.
* The rules of probability.
* Conditional probability.
* Bayes Theorem.
* Data distributions.

This puts you in a great place to tackle our next topic – hypothesis testing.