Module 3

Video Transcripts

Video 1: Introduction to module three

It is the core tenet of modern AI systems that, prediction and decision making is a probabilistic exercise. Application after application, practitioners in the fields of machine learning and AI have realised that in real-world applications it's better to think in probabilistic terms.

For example, what is the likely facial expression of this person on the image? What is the likely sentiment of this analyst's report? Conditioned on this market indices, what's the likelihood interest rates are going up by one percent? These questions are meant to illustrate that queries to AI systems are often probabilistic in nature, and one of the reasons why we design AI systems is in helping us quantify uncertainty in complex scenarios.

This module provides you with some foundational ideas in probability theory. So, you will be better equipped with the right language to navigate, build and discuss AI technologies. We give special attention to Bayesian probability, as well as using Simulation techniques such as Monte Carlo methods for statistical distributions, and to approximate quantities of interest in the real-world scenario. Join me in learning how to speak the logical language of AI.

Video 2: Introduction to probability theory

To help you appreciate the idea of probability, it helps you get up some perspective, some historic perspective of where probability is coming from. In the 16th century, Mathematicians and Gamblers captured the idea of chance, as in 'Games of Chance', and transform that idea into the discipline of probability. All these games involved coins, and dice, and cards. And so, these were the first models that people were interested in understanding, and those are the first models we're going to be discussing this module. So, let's get started and see some examples.

Video 3: Statistics vs probability

Today I would like to start by discussing some really, some basic and very interesting models from probability theory that are going to help us build some intuition for more elaborate models that we're going to see down the line.

So, in order to understand the difference between statistics and probability for this module, let's note them here. So, for us here, statistics will be the process of going from data and fitting a model that explains that data. Probability theory, on the other hand, at least in our context here will be concerned with, given a model, we're going to try to understand the data or the properties of the data that, that model can produce.

So, this is going to be our focus from now and later on, we're going to go through the inverse process. So, the first example that we're going to see in probability theory is that of a coin or coin flips. So, normally a coin has two sides. It has a head which is represented here by a face, and it has a tail which is normally represented by the crown. So, let's see you flip this coin bunch of times. So here, just for the sake of illustration I'm going to flip my coin four times, and this is the result that I got.

So, I have a question for you to get you thinking. Do you think head will come up more often than tail? Think about it for a second.

Video 4: The fair coin model

So, what's the answer that you have come up with? Here's my answer to you. It depends on what is the model that we are assuming that is generating that data. If we assume that we're dealing with a fair coin, then we should not expect more heads then tails. In fact, the frequency of heads and tails for fair coin should be approximately the same, or 0.5, or if you will, 50 per cent. This does not mean that we are going to get heads or tails in alternated ways. Sometimes we may get more heads as in the example above, sometimes we may get more tails.

But the idea is that, on average, and we're going to come back to that term again, we should expect a more or less balanced distribution of heads or tails as we flip this coin over and over again. So, now I have a second question for you. Is it possible to have a coin in which the distributions of heads and tails is unequal? Or said differently, is it possible to have a physical device, a coin or a die, for which you would expect certain numbers or certain realisations to come up more often?

Video 5: The loaded coin model

Now, I would like to discuss slightly variation of Fair coin model called a 'Loaded coin'. Difference between a fair coin, and a loaded coin is the relative frequency between heads and tails, which is no longer symmetric.

So, for example, I could have a loaded coin for which frequency of heads is 0.6, and therefore frequency of tails is 0.4. I could have a loaded coin for which the frequency of heads is 0.75, and the relative frequency of tails is 0.25. And, in a very extreme case, I could have a loaded coin with a relative frequency of heads which is 0.99 and the relative frequency of tail should be 0.01. In this last case, that means that if you were to flip such a coin, which sounds pretty much like a scam, I would obtain a sequence of heads nonstop for a long time. So, you would rarely see a tail.

In general, notice here that once again, I have a law for total probability. That means that the sum of the probability of the possible outcomes add up to one, as you would expect intuitively.

Video 6: Two flips of a coin

Next, we're going to consider a slightly more challenging problem, which is that of two flips of a coin. Let's say you flip two coins. So, I have coin one and then I have a second coin here which is coin two. So, if I flip two coins consecutively, what's the probability of getting heads (or H)? And the trick for solving this type of problem at least for now is, doing some sort of tabulation or building a table of possible outcomes. This works only for small problems, but it helps us build some intuition about probability.

Here, it is, how it goes. So, you built a little table here. On the first column, you're going to write down the outcomes for the first coin and for the second column write down the possible outcomes for the second coin.

So, I could have head and head, I could get here head and tail. I could have another tails and head or, I could have gotten tails and tails. So, the question could be interpreted in two ways, or I could be a bit more specific here. So, what is the probability of getting at least one head? Well, the first row gives me at least one head. And so does the second and third row. So, three out of four possibilities gives me at least one head.

Now if I were to ask myself a slightly different question, what is the probability of getting exactly two heads? Then in that case, the probability would be 1/4 because there's only one outcome that leads to that possibility. So, that's it for now. Two flips of a coin. I would like you to try next,

how you would compute different probabilities for three flips of a coin. Your table here is slightly bigger, but the type of reasoning, it is the same.

Video 7: Independent variables

Now that we've started discussing flips of multiple coins, let us look at an important property of random variables which is that of independence. Let's look back again at the occurrence table for the flips of two coins.

So, I had flip one and flip two. Then I could have head and head, head and tail, tail and head, head and head. Now, I'm going to represent this realisation given in a bit of an abstract way and I'm going to write this here as P(X_1) and P(X_2), P (X_1, X_2), where X_1 can take a value of heads or tail and X_2 can also take values of heads or tails.

What you can verify yourself by looking at this table here is that the probability that X_1 takes the value say H or X_2 takes another value T that is equals to the probability of X_1 and the probability of X_2, and you can check that. So, for example, if I have here probability of X_1 it was head and the probability of X_2 equals to head that is equivalent to P(X_1) equals eight times of probability of X2 equals H, and in probability parlance this is as saying that this means that these variables are independent.

So, in general when I flip coins in the number of coins, the results of those flips, intuitively speaking, they're independent of each other. And the way to represent that mathematically by saying that when events are independent, the probability of them co-occurring or occurring at the same time is the product of those probabilities. This relation here does not always hold true, but when it does, it is a very strong and important property for us. So, you can actually generalise that, like I said, for any number of coins, and in probability theory that's called the probability of a composite event.

Can you think how this formula would generalise if instead of two random variables, say two coins, I had three coins?

Video 8: Probability theory: summary

I hope that by now, you are able to model and compute the probabilities for simple discrete events like rows of dice or sampling cards from a deck. And that you now understand the concept of sequences of independent events. Stay tuned for the next lesson.

Video 9: Introduction to Bayes' rule

This week, we're going to talk about Bayes' Rule. Bayes' Rule is a cornerstone of modern statistics, and it is the mathematical foundation for something called, 'Statistical Inference'. Statistical inference means that, with Bayes' Rule we can find the predictive probability, the probability of something happening, based only on our understanding or assumptions of what happened in the past. This is a slightly different from the frequency approach we took before, where we estimated probabilities based on the frequency at which things happened, like tosses of a coin or throws of a dice.

Bayes' Rule, if you don't know, it was also fundamental in the war effort. And here in the UK, Alan Turing and his group applied it to break codes too, when they were fighting against the Axis. So, let's learn more about the Bayes' Rule.

Video 10: Bayes' rule

So, now that we have some intuition about probabilities, let's talk about a very important and core concept in probability theory, which is that of a conditional probability. And I'm going to try to make a drawing here to give you some intuition. For this explanation, I'm thinking of probabilities in terms of areas. The square here is the space of all possible variables, of all possible occurrences of say a variables, X1 and X2. And this is sort of like this area here, red area, this is equal to the probability that a variable X1, has taken some value.

Okay. Which we can call here little X1, then we have your second area here in the second area here, blue area. The probability that a second random variable X2, It's taking a second variable X2. Okay. And what I want to define, and I think the picture really helps here, keep an eye on this intersection here, right? Which is what we call a conditional probability. So, we do not. So, conditional probability. And the notation we use is what is the probability that X2, took some value? Say X, little X2, given that X1 has taken some value X1. And we're going to define it here, and then you get, you know, understand the intuition for why this is the probability that X1 and X2 happen so the intersection of those two probabilities. But compare to the probability or a fraction with respect to the probability of X2, given that X1 happened, what is the probability that X2 happened? And that conditional probability here is the fraction of this little area here. This intersection here in relation to the bigger area here. So, dramatically that's what conditional probability is. So, this is sort of like a geometric intuition.

Now, given this geometric intuition, a very important consequence of this fact of conditional probability, we have also here P of X2 given X1. If you do a little bit of a permutation here of the variables you will obtain that P(X1) given X2 is equal to P(X1) given X2, divided by P(X2). And the curious fact is by combining this two formulas here, what you obtain is what we call, Bayes' rule. So, mathematically seems like a trivial fact that P(X2) that P(X2) given P(X1) is equal to P(X1), given X2 times P(X2), divided by P(X1). But the fact is this has some deep philosophical consequences that we're going to talk about in the next videos. This is a core idea for machine learning and there is some important interpretations that come with it. So, stay tuned.

Video 11: Total probability

The reason we started talking about Conditional Probabilities is that they are extremely useful in when we're trying to express certain beliefs about events. So, let me give you a concrete example. I'm going to talk about a medical example, and let's assume that there's some rare disease for which, in some population, imagine specific bracket of demographics, probability that someone is ill is 0.1. And the probability that someone is not ill, which I'm using this symbol here, you can also use a little squiggle, is the complementary probability, which is 0.9. Now imagine that you have some test for this rare disease we're talking about, and for that disease we have the following data. That the probability that the test is positive, given that you're ill, is 0.9.

And I have further that the probability that the test gives you positive, given that I'm not ill, so it's a false positive, it's 0.2. And the probability that I have a negative result, given that I'm not ill is 0.8. So, notice that if I'm ill or not ill, these probabilities are complementary or in other words, they both add up here to one. So, this is very, this comes in a very interesting way, because by combining conditional probabilities with absolute probabilities, we have an extra power, we have an extra tool for expressing certain probabilities.

So, for example, let's say I'm interested in computing the probability that I'm going to get a positive result. And the probability that I'm going to get a positive result is, probability that I'm going to get a positive result, given that I am ill times the probability that I am ill, plus the probability that I got a positive result, given that I'm not ill times the probability that I'm not ill. And this formula is in a way consequence of Bayes' Theorem. But, depending on where you start in probability theory, this can actually, could be stated as a axiom, or a basic fact of the

probability system.

So, just to recap here, I can express this total probability here in terms of other probabilities, other partial probabilities, and probabilities that we've talked about. So, I can actually compute what is the probability of getting a positive result by combining some of the elements that we've seen before. So, bear this formula in mind because it will come in very handy again and again in some of our later exercises and examples.

Video 12: Applications of the total probability formula

So, we have talked about the total probability formula and I'm going just to recap it here briefly, which is just of expressing a total probability in terms of conditional probabilities. Here for now, I am using a simplified version of the formula, assuming that my variables are binary or Boolean variables, and they can only take two values, binary like in the flip of a coin. This is then called the total probability formula. So, let's see some examples here and see how we can use this formula to actually solve some fairly complicated problems. The first one I want to consider is: Imagine you have two coins, one is a loaded coin, one is a fair coin.

So, I can get heads and tails with a 50 percent chance. And then I have a second coin which is loaded, and I'm going to get heads and tails with different probabilities. I'm going to get heads with the probability of 0.9. So, now imagine that you blindly choose one of these coins, and I'm assuming here you have 50-50 chance of choosing either of them if you're choosing them blindly. And you flip whichever coin you've chosen twice, and you end up with heads and tail, in this order. So, what is the probability that this could have happened, given that you don't know which coin you've picked up? So, one way to go about this problem is by using total probabilities, and we conditioned it on a second random variable which is the coin we've chosen.

So, for example, this total probability could have been expressed as total probability of getting heads and tails, given that you've chosen the first coin, times the probability of getting that first coin, plus the probability of getting heads and tails exactly, given that you've chosen the second coin, times the probability of getting the second coin. Now, once you've chosen a coin, you can compute what is the probability of getting heads and tails. So, for the first coin that's 0.25, because it's the probability of getting heads, which is point half, and then probability of getting tails, which is point half again, this time 0.5 because I have a 50 percent chance of choosing the first coin. Plus, here for the loaded coin, the probability of getting heads and tails exactly is actually a lot smaller. It's 0.09, because I'm actually more biased towards getting a head.

When I multiply this by 0.5, if I add this all up, I'm going to end up with 0.17. Okay. Now, I have a slightly more challenging problem for you. Imagine now that you have the possibility of changing coins. So, my second example here is: Once again I have the two coins. I have a fair coin here, heads and tails, with my odds here being fair.

I have my loaded coin here, heads and tails with a heavier bias towards heads, of 0.9. So, once again I flipped these coins twice, and I got exactly heads and tails. Now, with the caveat that I can change coins between the first flip and the second flips. So, in this case, what is the probability of getting exactly heads and tails? One way of solving this problem is, once again using the total probability formula. Here that's how we're going to express it. So, the probability of getting heads and tails, but now we're going to condition it on a variable, which is a variable that expresses our choices of coins.

Then there is four possible choices of coins. We can either choose the first coin twice, or we can use alternate versions of those coins, or we can use the second coins twice. I'm going to leave it to you to think a little bit and actually compute those numbers. So, the formula will look like this, given that I end up with here one and one. So, this means I got heads and tails, and in both cases I had the first coin. So, what is the probability here then that I choose coins one and one, plus probability of heads and tails, given that I choose first coin number one and then

coin number two, then I'm going to get here probability of heads and tails, given that I have two and one, probability of two and one, plus finally here the probability that I get heads and tails, given that I get coins two and two here, probability of two and two.

And I will leave it as an exercise for you to compute this probability, and in the end, you're going to end up with 0.21, which is bigger than our previous odd of 0.17. So, in other words, if I'm allowed to change coins between flips, my odds improve here by a factor of 4 percent. Keep that in mind if you're betting money against someone. Let's see next, Bayes' theorem and applications of Bayes' theorem.

Video 13: Bayes' rule applications

When we talked about Conditional Probabilities, we have also introduced Bayes' Rule and Bayes' Rule was a way of inverting probabilities which I will explain in better detail in a second. So algebraically, Bayes' Rule said something like this; that the probability of A given B, once again for the sake of concreteness, let's imagine that is our binary variables. The probabilities of A given B is the probability of B given A times the probability of A divided by the probability of B. In other words, I can express this probability here in terms of this probability here.

And originally, this relationship was called an Inverse Probability and it was one of the first instances of inference using probability theory that we are aware of. So, let's see how that works. Imagine again that we have some rare disease for which we know that the probability that you're ill with that disease is 0.01 in our population and imagine that you have a test for that disease. So, imagine here they have, for example, that you are 90 percent confident that this test is positive if someone is ill and I am 90 percent confident too that the test is negative if someone is not ill.

So, the big question that we want to ask ourselves is, what is the actual probability that I have the disease if I get positive test? So once again, I have expressed here the probabilities of getting a positive test if I'm ill. But how confident am I that I'm ill if I get a positive test? So, this is an example of an inverse probability problem. So, let's see what we can do here, and this is where we're going to apply Bayes' theorem. So, according to Bayes' theorem, the probability that I'm ill given that I got a positive result should be equal to the probability that I got a positive result given that I'm ill, or someone else is ill, times the probability of someone being ill divided by the probability of a positive test.

And now we're going to use a second tool which is, we will use the total probability formula here and I'm going to re-express this term here in the denominator as sum of probabilities that I have already knowledge of. So, for example, I can re-express this as probability of getting a positive result given that I know that the person is ill times the probability that someone is ill divided then by the probability that someone got a positive test given that they are ill divided by the probability that they are ill plus probability of getting a positive test given that someone is not ill times the probability of someone not being ill. And I would like you to, as an exercise, to compute some of these probabilities here that we haven't directly expressed. So, for example, this probability here and this probability here. So, you can actually finish this exercise and with a little bit of patience in algebra here, if you have thought about it a little bit, I'm going to get the following result. So, this is 0.9 times 0.01, divided by 0.9 times 0.01 plus probability of getting a positive result if you're not ill, you can check it yourself, is 0.1 and the probability of not being ill is 0.99. And just to keep your brain warm, I want you to finish this algebra here and cook through the final answer. So, to recap, we have managed to use information about our confidence on whether a test succeeds or not to then determine the inverse probability of how confident we are that someone is sick given that we've got a positive result. It is amazing how with a little bit of algebra and logic, you can solve a fairly complicated problem.

Video 14: Bayes' rule summary

I hope you had fun this week learning about the Bayes' Rule. By now, you should be able to frame and compute simple problems off the form given that A happened, how likely it is that B happen too and vice-versa. This is a big conceptual leap because often times if we don't have much evidence or enough evidence about something, we still have to make an intelligent guess and update our belief based on incoming data. It's a big and powerful idea of modern data science and machine learning. I hope you appreciate it. Let's move on.

Video 15: Monte Carlo simulations

We will introduce a broad class of computational algorithms, also known as Monte Carlo, that rely on random samplings to obtain approximate numerical results. Fun fact, Monte Carlo methods were invented when a scientist, Stanislaw Ulam was playing solitaire and wanted to figure out his odds of winning. Monte Carlo simulation and it's one of those ideas so pervasive to numerical and computational probability and statistics these days that a lot of times we are using it without even being aware of what it is and how it's done. And to be very honest, there is not quite a definite definition as to what Monte Carlo simulations are.

Historically, they came to be during the war effort in the '40s at Los Alamos by two or three prominent scientists there, including Stanislaw Ulam and John von Neumann. And really what they wanted is to be able to use a computer to simulate random processes at the time they were using it for chain reactions in nuclear physics. But it came to be such a versatile tool that's used all over the place, finance included. So, what I want to show you here is to give you a little taster of how Monte-Carlo simulations are used in practice in many different sort of like stages of a statistical or probabilistic computation.

So, what we're going to start by doing here is to compute we're going to try to simulate a random walk, right? A lot of times this is called the drunkards walk, right? Because you imagine you have a drunkard, leaning onto a lamppost and he can take a walk to the left or to the right, and he can end up pretty much anywhere. And the way you do it, it's almost like simulating a coin flip, but instead of the random variable, which is the walk outputting a plus one or a minus one, one or zero, which is often the case used in other types of simulations, we use here one and minus one because we are interested in the final position, say, of the final position calculated by the random walk.

So, for example, when I actually simulate a random walk, I get a sequence of zeroes and ones, but the actual random walk is going to be adding up those ones and minus ones. And you can see how useful that is, for example, in the stock market, especially for people who are interested in, say, simulating or trying to determine pricing for stock options. We're not going to delve into that here, but just to give you an idea of like some of the applications, okay. And for that particular around the walk there, what happened was when I calculated the average is 0.18.

It's a little skewed to the right, and that's by design because I actually had set that 60 percent of the time I should be getting, I step to the right and what we're going to do here next is to simulate a random walk. And basically, it's just building on top of what we have done before. So, we're going to actually build lots of the simulations, right? We're going to actually have them stacked up into a large frame. And each role of these large matrix here is going to represent one path. And important to notice that we're actually taking the cumulative sum of those simulated plus or minus ones, so that you actually can get sort of like a definite position, towers the end of those 100 steps.

And then I run a simulation. So, I have paths of size of 100 and I get a 100 paths. So, this array here alone has 10,000 entries and you can see how these things, they get very quickly, very large. And that's why Monte-Carlo simulation is a computer intensive technique. And then when I actually ask python and the computer to plot that for me, I get these spaghetti plot here

of all the different random walks in different colours. But you see how interesting it is that even though we always start here at point zero; some walks, they're going to stay around zero, some walks are going to diverge all the way to 20 or to -20.

Okay. And you can actually verify that, the average position at time 100, let's say after 100 simulation steps, is near zero. And standard deviation, for one standard deviation is about 10 steps. And that kind of matches with sort of like the visual intuition that we get from just looking at this plot. Traditionally, another reason for using Monte-Carlo simulations is in the computation of areas or volumes or all sorts of volume related quantities, which often in mathematics and statistics, they come about in the form of what's called integrals. And if you're rusty in your differential and integral calculus, not to worry.

The idea is you can use Monte-Carlo simulations for estimating the volumes of things and the way you do it here, what I'm trying to do is the following, I'm going to try to estimate the value of pi, the constant pi that shows up in mathematics when you're trying to compute for example, the circumference of a circle or the area of a disk.

And the way I do it here is, I simulate two random uniform variables. So, it's within a square of side one. Okay. And so that means that my circle has radius 0.5 and I sampled that pair of points, right? And if they fall within the disk, that means that I'll count plus one, right? So, I only count the points that fall within the disk. So, imagine that, like I'm throwing darts, into the square screen and the ones that fall within the circle, I count them. And to actually determine the value of pi is you actually estimate the size of the area of the disk by counting the number of points that has fallen within the disk.

And that's why, I'm actually dividing count by the total number of points that we used in the simulation. And then if you're further divide that by 0.5 to the second power of 0.5 squared, I actually get the value of pi. If you remember from basic, from school geometry, area of a disk is equal to pi times the radius of the circle square. So, if you divide the area by the radius square, you get the value of pi. And what happens when I actually run the simulation, I get a value here, which is not too bad.

It's 3.11, when the value of pi is 3.14. So, here it is, this is Monte Carlo Simulation. It's this pervasive. It's sort of like a Swiss knife type of tool for computation in statistics and probability that will come up all over the place when you study statistics, statistical simulations, and all sorts of things related to AI, especially when you're trying to simulate agents and things like that. So, that's all for today.

Video 16: Probability distribution: introduction

The basic question probability distributions help us address is: What is the probability of an outcome of a random variable? Data, for example, is the realisation of a random variable. So, when I get a head or a tail in a coin toss, that's the realisation of a random variable which is modelling the coin. If my random variable has a discrete number of outcomes, a simple table will do. For example, if I throw 10 coins and add the number of heads, I can compile a table with all the possible results. The simple experiment of counting the number of coins in a toss of multiple coins will lead us to the Central Limit Theorem, which in turn lead us to asking yourselves; How to model the probability distribution when the number of outcomes is too large, let's say, potentially infinite.

Also, in practice, a lot of random variables are not discrete. Imagine for example, people's height, weight, body temperature; there are not discrete increments there. In this case, we need a function representation of the probability distribution like a constant function or a linear function or even exponential function. One of the most common continuous probability distributions is the normal or Gaussian Distribution, which we will learn how to use this week. It will extend your toolbox for solving really interesting and cool problems in probability, so stay tuned.

Video 17: Binomial distribution: part 1

We've talked about flipping coins before, and as a motivation for what's coming next in our statistics module, I would like to introduce an important concept from probability theory, which is that of a Binomial Distribution, which will lead to another important distribution, called the Normal Distribution. And here, we'll see a distinction between discrete and continuous variables. And, what's important here is that the binomial distribution comes about when we start counting and taking averages. So, let me motivate the type of problems that we're going to be dealing with from now on. So, let's say that I have a coin as usual.

It's a fair coin to begin with. So, heads and tails 0.5 chance of getting either. And I flip this coin five times. Okay. What's the probability that I get exactly one head in five tosses? Think about it for a second. If you have thought about it, now we need to introduce combinatorics, or we have to review some notions from combinatorics. There is about a couple of different ways on how to solve this problem, by one of them is by counting. Okay. So, when I flip a coin five times, there are five different ways in which I can get exactly one head.

So, if you imagine here, five slots representing the five flips of a coin, then head could have come on the first slot, or the second slot, or the third slot so on so forth. In five flips of a coin, there is about 32 possible outcomes. And, by dividing these two numbers here, we end up with approximately 0.16 or 0.15625. Now, things start getting a little bit more complicated, if you ask yourself a slightly different question which is, probability of getting exactly three heads; what do you think that answer should be now? And, for that, we're going to need a new formula to help us with this, which is the so called Binomial Coefficient.

So, the binomial coefficient, I can write it like this, and I'm going to define it here informally as, what are all the possible ways of getting, let's say 'k' heads in 'n' tosses, and the formula that you get, is something involving factorials, and the formula looks like this. It may look a little complicated to begin with, but it's actually fairly simple. So, if we go back to our original problem, the probability of getting exactly three heads in five tosses, is going to be equal to, the probability of getting, which is going to be 10 over 32, which is approximately 0.3125, and I'm going to ask you, to try to derive that number by yourself. Once again, by using the binomial coefficient formula, there are 10 possible ways, in which I can get three heads in five tosses, and the total number of results that I can get after five tosses is 32. So, this relative frequency is the probability of me, getting exactly three heads in five tosses. Okay. As an exercise, for you to keep going, I would like you to compute, the probability of getting exactly one head, given that you have a loaded coin with a bias of 0.9. What do you think that could be?

Video 18: Binomial distribution: part 2

So, next I would like to do a quick recap of the binomial distribution. So, first let me talk about Binomial coefficients, then Probabilities. Then I'm going to talk about Histogram of frequencies and how that's related to a probability distribution. Binomial coefficients, you can generate down by a procedure called Pascal's triangle. And it works like this. You start here with the one then you have here one and one. And the way you start filling this triangle here is sort of like adding up adjacent neighbours upstairs. So, you imagine there's a zero and a zero here.

You imagine there's a zero and zero here, so on and so forth. So, there's a zero here in every single stack of this pyramid. So, for example, the next one here is also a one. The middle one here is a two. This one here is also a one. This one is also going to be one. This is going to be a three. This is going to be a three, this is going to be a one, then the next one here is a one. A four, a six, a four, and a one, so on so forth. And one way of interpreting each one of these coefficients is by that.

Let's see here, imagine that each line here represents the number of tosses. So, for example, this line here, I flip a coin twice. So, this is the number of times I can get exactly no heads, this is the number of times I get exactly one head and this is the number of times that I can get

exactly two heads. So, this subsequent line is flip thrice or three times. So, this is the number of times I got no head. This is the number of times that I get exactly one head. This is the number of times that I get exactly two heads and this is the number of times that I get exactly three heads, so on and so forth.

So, you have this recursive way of producing the combinatorics that tells you what are the many possible ways that I can get a certain number of coins, a certain count of heads. And this is important because if you add up the score efficient, if you add up all the rows, you end up with all the possible outcomes of say here, two tosses of a coin or three tosses of the coin. And the next one here is four tosses of a coin. So, that's 16 and you can check that the subsequent rows, they will all add up to powers of two and therefore that's why we end up with a formula that looks exactly like this, that the number of heads equals 'k' in 'n' tosses.

This is equal to the corresponding entry here in the Pascal triangle, which we denote by this number here divided by two to the power n and in a nutshell, this is the binomial distribution, which is the distribution that represents the counts or the sum of heads in a sequence of k heads. It's this distribution that represents the sum of heads in a sequence of n tosses. Now there is a slightly more geometric way of representing these distributions in the form of a histogram, and that will lead us next to the central limit theorem and the normal distribution. So, stay tuned to hear about the rest of the story.

Video 19: The central limit theorem

In this presentation, we're going to be talking about the Central Limit Theorem, which is a fundamental result from the probability theory that is applicable and useful in various statistical and modelling problems. And here it is how it goes. Let's say that you are counting the number of heads after a sequence of 'n' coin tosses. You count the number of heads. We've seen before that for a coin, which we're going to represent here as a simple parameter or a fair coin, and I'm going to have sigma different from a half, for a loaded coin. We've seen before that, if you were to estimate the probability that we get exactly K heads say, then this probability is equal to $\mathbf{C_k^n}$, theta to the k, one minus theta to the n minus k.

This is all nice and dandy if n is a small. However, as n gets larger and larger, there's no number here, which is the binomial coefficient gets increasingly harder and more cumbersome to compute. And the reason for that being, it involves factorials, and factorials of large numbers are computationally intense. So, what people observed and if you remember the histogram plot that we had for the binomial coefficients, what happens is, as n increases, I get histogram plot. Let's see here, frequency, in terms of K that starts picking up and starts getting increasingly concentrated here in the middle.

So, this peak here is going to be exactly the binomial coefficient. And the middle here, and it's peeking at the floor function of n over two. So, for example, if n where say a 100. And n over two for is 50. So, that number there would be 51. What people observed, or the great minds of the past, is that after the appropriate re-scaling in the limit this empirical histogram, will approximate a Gaussian curve with mean n theta and with variance square equals to n theta times theta minus one. This is useful because I can then say that my function 'f' above, which is counting the number of heads in a sequence of n tosses, gets approximated by a normal function of say mu, n theta and variance n theta times theta minus one.

Video 20: Applications of the central limit theorem

In our previous presentation, we've introduced the Central Limit Theorem. Let's talk about some applications of the central limit theorem now. So, we've started with Coin Tosses, then we moved on, earlier in this module, into the Binomial Distribution and our next step was to connect this to a normal and or Gaussian Distribution. And the connection here from the binomial distribution to the normal distribution is the so-called Central limit theorem (CLT).

More abstractly, if we call X a random variable which is a sum of any Boolean variables, meaning each one of these variables can either be a zero or one, you can think of them as abstract representations of coins. So, each one of these variables here represents coins with a certain parameter Theta, and what the central limit theorem is telling us is that this random variable here has a probability distribution, which I'm going to call here P(x) which is normal with mean, n Theta and with variance squared n times Theta times one minus Theta. And why is that useful to us? In a more general sense, this theorem not only applies just sums of Boolean variables but sums of random variables, independent and identically distributed.

This is called the Laplace-De Moivre theorem for Boolean or Bernoulli variables, but this theorem can be a little more general than that. So, let's see an application here, let's see an example here. Let's see, we have n equals to a 100, then we have here Theta equal to 0.5, which is a fair coin. So then in that case μ is going to be 50 and Σ square is going to be 25. So then, in order for me to estimate the probability that, for example, I would be able to achieve probability of, say k equals 10 heads which you could easily compute by consulting a table or specialised software.

Video 21: Probability theory: manipulating normal variables

Since we've got into the normal distribution, coming from the binomial distribution. Quaint authors let's talk about some interesting properties of the normal distribution, that will help us gain some intuition about how to manipulate them. So, let's say here, I have a bell-shaped curve describing the salary distribution at Imperial College. Let's say here average here is 50K and that my dispersion here is around 20K. So, this is my distribution of salaries at Imperial and I'm assuming it to be a normal distribution. Let's say, for example, that the provost decides to give everyone a 10K raise.

How do you think the mean and the standard deviation will change as a consequence? So, my mean is 50K and my standard deviation here is 20K. If you thought about it for a second, then you will realise that my mean will shift, also by 10K. But the dispersion, the standard deviation, will remain the same. And geometrically you can see that because by shifting the mean, what I'm doing here, by giving everyone a 10K raise, I'm just shifting the mean. But the spreading salaries won't change. Now let me ask you a different question, let's say the provost got really happy with everyone's performance, he decided that he will instead of giving everyone a 10K raise, he will double their salaries.

So, this is part one and this is part two, what do you think now it's going to happen with Mu, the mean, the new mean and sigma prime, which is the new sigma. Think about it for a second. So, if you thought about it for a second, if everyone's salary double, the mean will also double. But what's interesting is that the standard deviation, will double as well. So, that means that our dispersion from the mean will double, and that makes sense because small salaries won't change much, but big salaries will change quite a lot. So, you can see that, that curve will somehow, flatten a little bit.

Now, next, I would like us to discuss some interesting properties about how uncertainty compounds using normal curves or using random variables, which have normal distributions. So, let's say you are a golf player here, with your baseball hat and you are at your favourite golf club, you hit your ball, and you know that on average your drive is about a 100 meters with a standard deviation of plus or minus five meters. So, if you hit, if you drive a golf ball twice, how does the mean change to mu prime? And how does the standard deviation change? What's interesting in this case, if you thought about it is that, means, they always add up or they could also subtract but standard deviations, they will always add up.

And I guess an intuitive way of seeing that, is that imagine that you've driven your golf ball twice, then your range now will be, say 200 meters. But my uncertainty compounded. And now, I have a 10 meters uncertainty of where the ball will land. So, I'm assuming here I'm driving the ball in the same direction. Let's say, instead that you drive, do a second thought experiment

and imagine that you drive the ball instead in opposite directions. Then on average, the ball will be on the same place, but with an uncertainty again of plus or minus 10 meters, so average is cancelled out.

But uncertainty or dispersion compounded. And I would like you to keep these rules in mind about how uncertainty compounds for normal distributions. I'm assuming here that my golf drive is normally distributed with an average of a 100 meters and a standard deviation of dispersion of five meters. And the second set of rules that I find useful to keep in mind is, how my mean and my standard deviations, they scale when I scaled. As my example, that was the distribution of salaries at Imperial College. I hope those new tricks will be useful for you when we start solving some very interesting problems later on.

Video 22: Summarising module three

In this module, we have learned about various Probability models and started to understand how they related to key ideas in Machine Learning. We learned these new ideas conceptually first, and then as a next step, we designed numerical experiments in Python to understand how they work in practice. Along the way, we introduced a few new libraries from Python including NumPy, Pandas, and SciPy, which enabled us to perform computations in a painless way. May this be the first of your many excursions into probability theory.