Module 9
Video Transcripts

Video 1: Uniform Distribution (6:40)

Hello, and welcome to module nine. In this module, we'll cover different statistical concepts and tools that will help you understand better your data. In the last module, you learned about topics like distributions, random variables, probabilities, and all of this, and now we're going to learn how we can apply all of that to data science. The thing we're going to do is in this first network we're going to cover different statistical distributions. We're going to understand the most important ones that are the uniform, the Bernoulli, binomial, normal, exponential, Poisson, and the t distribution. So, let's start by importing all our libraries for plotting and creating data that are numpy and seaborn. So, the first distribution we're going to understand is the uniform distribution. This distribution is also known as the rectangular distribution and is the one that has a constant probability. The probability distribution function of continuous uniform distribution is this one here. So, it's a piecewise function, as you can see, that it's 0 if x is less than a, our value a, 1 over b minus a if x is between a and b, or 0 if x is bigger than b. We will be using the scipy for the rest of this notebook, and the idea here is that you can create these functions by yourself. I mean all of them have, they all have an equation, so you can easily with Python program that equation, but it's much better and easier to do it with good libraries like scipy.

One other advantage of using scipy like numpy is that it has some optimization underneath. So even though you can create code that is similar to this, it's very unlikely that you will reach the performance of these libraries. So, to begin, let's import the uniform distribution from scipy. We're actually using a scipy module called stats. So, in able to do all of this, the first thing we need to do is to create random numbers that will come from a uniform distribution. The idea of this in this module we're going to learn how to create the distributions knowing that we know, I mean, with the knowledge that we are aware that we have all the distributions. The idea is that in real life no one is going to tell you this data comes from this unless you have a very specific process. But what you will do is that you will understand your data, and you will plot it and perform operations, and you have to discover if this data is coming from a normal distribution, binomial, and all of that. So, with that in mind, let's start by creating 10,000 random numbers from the uniform distribution, and what this function takes is the start and it takes the scale. To do that, we're going to use the function rvs. I'm going to create this, and then I'm going to show you what's rvs. Rvs means random variates of given type. This is the way we can create an array of different values for a given distribution. So, it's just creating random numbers that comes from a distribution. These arguments here are not the same for each one of the distributions we're going to see in this module. As you will see, this right here needs a size, a loc, and a scale, but not all of them are the same. So, now we created our data uniform. Let's first try to see it. As I told you, there are numbers between 1 and 20, and there are 10,000 of them. They can be, I mean in the uniform distribution, you don't need to have only integers. You can also have decimal values. The important thing here is that you need to follow a uniform distribution. The way we're going to check this is a distribution plot. In seaborn it's very simple to plot a distribution. What we'll have to do is to use the function distplot. Then, we're going to pass this distribution the data uniform that contains the data from the uniform distribution. We're going to say bins 10, 20, sorry, and kde false.

This distribution plot, if you say kde is equal to true, it will also print kernel function here that will follow along the path of the distribution, but it's not in the bars what is actually a line, a smooth line. If you want to see it, you can just say kde equals to true, and here you'll see it. You can take the same amount of information from this line than from the bars. That's very important. And it's interesting here, the distribution function should look like a rectangle, something like this, but it's not quite a rectangle.

The problem here is that we're not using enough data to be able to account for this type of distribution. So, 10,000 points may seem like a lot, but you need more, much more points, to be able to get into the actual shape. A perfect distribution will have an infinite amount of numbers, that's not possible, but we can grow, and we will see that it's going to get much more closer. Just for the sake of trying, let me go here and just test for 100,000 points. Now, I'm going to plot it here, and as you can see, it's much more close to the actual data, to the actual form or shape we expected from the uniform distribution. If I go to a million points, 10 million points, 100 million points, a bill points, it's going to get closer and closer and closer to the shape of the rectangle.

****

Video 2: Bernoulli and Binomial Distributions (6:25)

Let's go now to the Bernoulli distribution. By the way, the people always hear the name Bernoulli. It's not always the same guy, that's an interesting fact. You can check, they were a family of mathematicians, physicians, and doctors, and you can see the history of the Bernoulli family. They're very interesting people that added a lot of value to the world of science. So, coming back to the world of statistics. The Bernoulli distribution has only two possible outcomes, the success and the failure. Normally, we say the success is 1 or the 0 is the failure. Like, if we say we toss a coin, we can say that heads is success, so 1, and that tails is 0, so a failure. The function we have for the Bernoulli distribution is this one right here. So, it depends on a value called k, and also, and this k is 0, 1, here, the success and failure, and this p is the probability of success. So, it depends only on two parameters. We have k and we have p, where k is 0, 1, and is the probability of success.

This probability of success shouldn't always be 0.5. That means that we're used to understanding like the, when you toss a coin, you will get a 50/50 percent of getting a head or getting a tails, but you can also have a loaded coin, and with a loaded coin, you'll find that sometimes the probability of heads is bigger than tails or backwards. For the sake of the example here, we're going to use a Bernoulli distribution with p equals to 0.5. In here, k is assumed to be 0 or 1, and we're going to create 10,000 cases again. So, again, probability is 0.5. You can see we're also using the stats module in scipy, but now we're importing the Bernoulli distribution. The rvs here, so the random variables we're creating is not the same as before. If you remember, in the uniform distribution, it asks us for the width, for the scale, something we called the loc, and all of that. In here, we're only asked to give the size and the p. So, when you do this, let me now run this, we have an array of 10,000 points that are always 1 or 0. If we plot it, as you will expect, this is what we will see. We will see a graph that will have half of the points in 0 and half of the points in 1. This is because the probability we gave was 0.5. If we say 0.6, as you can imagine, we'll have more data towards the 1. If we say 0.4, we'll have more data towards the 0. So, you can play with this and see different cases for the Bernoulli distribution.

Another distribution that is similar to the Bernoulli, but in here we have different trials, is called the binomial distribution. Binomial distribution is a very interesting one, and as you can see here, the term here on the right is almost the same as the Bernoulli distribution or that we have another factor and another term that is called the binomial here. And this binomial here, it's dependent on n and k. So, k is, again, 0 or 1 in this case, but n here is the number of trials we give. And one important thing is that each trial is independent of each other. But the definition of this symbol here, nk, is this, is a factorial of n divided by k factorial and that's multiplied by n minus k. So, again, you can create a function and the same with the Bernoulli distribution, but we're going to trust again scipy in Python. In this case, the way we use it is that we import binom here, the binomial distribution from scipy stats, and the only difference we have when creating the variate is that we have to give the number of trials. Here, we have 10 different trials of let's say toss in a loaded coin 10,000 times where the probability of heads is 0.6. You can think of it as like that.

Let me run this, and in here you will have, now you won't have 0 or 1. What you'll have here is something like this. It says here that, so 10 in this case is the number of trials you did. So, if you have a 6 here, this means that when you in the first trial you launched the, let's say we have a coin, you tossed a coin 10 times, and you got six heads. The next one you got 10, then 9, then 3, then 8, and then you have this distribution. If we plot it here, it's going to be very similar to a different one that is called the normal distribution, and this is the way we can plot it. So, if you remember some exercises and videos about like tossing coins and the central limit theorem, some of these things are very related to the concept of the binomial distribution, so this is a way we can start understanding all of that. This distribution is a distribution that is not a continuous function, okay. In this case, we don't have a continuous function. We have a summation of functions. That is not the same as having like a smooth function in this case.

****

Video 3: Normal Distribution (2:55)

The next distribution we're going to talk about is the normal distribution. The normal distribution may be the most important one in all of statistics. If you ask a statistician what you really need to understand to learn and understand statistics, they're going to tell you that you need to fully understand the normal distribution. You will see much more about it in the future videos, but for now you can just think of it as an important distribution that is followed by a lot of processes in the world of data science, which you learn in that. A lot of people call them, call this distribution the normal distribution. Some people, like physicists, they call it the Gaussian distribution. And there are some people that also call it the bell curve, because of the shape, we're going to see it very soon. The function behind the normal distribution is a distribution function density curve with mean mu centered deviation sigma. So, these are the parameters of our function. So, we have our x, our variable, and we have two parameters. We have sigma squared, that we also call the variance. If we take this 2 here, we have the standard deviation, and we have mu, that is also called the mean. And it's an exponential function, as you can see here, but it's a very specific type of exponential function. The way we create it here with Python is that we import it from Python, as usual, but in this case, it will ask us for three things. The size is how many points we want to create. So, in here I'm going to create 10,000 points. Loc is going to be mu, so we're going to create a mu 0, and scale is going to be 1.

And scale, in this case, is the sigma, okay? Not sigma squared, but sigma. In the case, sigma squared, and sigma is the same because there is 1, but it's that's not the case all the time. Something important here, when you have a normal distribution, that mu is 0 and sigma is 1, we call this this the standard normal distribution. So, let me create this and then, as you can see here, we have different points. And so, when we do this and we plot it, this is what we're going to see. You may have seen this before. It is the one they call the bell curve. It's a bell shape where almost all of the data is centered towards the mean. So, most of the data is close to 0 here, and it's getting, and we have less and less data if you go to the extremes of the graphic.

****

Video 4: Exponential, Poisson, and T Distributions (7:23)

For our final video in this notebook, we're going to talk about three more distributions. The first one is the exponential distribution. This describes events in a Poisson process. When you hear this, don't worry, it's not this weird thing. It's just a process in which events occur continuously and independently at a constant average rate. That's the definition of a Poisson point. By the way, it's not poison, it's Poisson, it's French. So, this was also a guy like Bernoulli. The general formula for the density function of the exponential distribution is this one here, and it looks like 1 over beta where beta we call it is scale parameter, and it has mu, that is also the mean. And so, that's basically it. But normally you'll see it online as this equation here. And when we do this, we call lambda the division between 1 over beta. So, normally we use this equation to express an exponential distribution. So, this is very simple. We just have to important it from scipy, and it's going to ask us for the scale, and the loc is going to be here, the mu is going to be the mean, and the scale here is going to be this parameter, lambda here and size. So, that's going to do this, this is what it's going to look like. And as you might imagine, it looks like an exponential function. But it's also, I mean you can also think of this as it's not that different from a normal distribution, meaning that you have this kind of a bell stuff here, but it's centered towards a different place. And it's not true in this case that all of the cases are very close to the mean, because the mean is 0 here, but we have a lot of cases also in the right side. So, this is the way we can represent the exponential distribution.

Let's go now to a Poisson distribution, and as you may imagine, this is very similar to the exponential distribution, but in here, we are modelling the number of events occurring in a given time interval. The math function is very similar to the last one, let me remind you here, we have lambda exponential minus lambda x, but in here, we have something like may remind you to the binomial because we have factorial and we have some divisions. So, this is the way we can put it here. By the way, with a little bit of math fun, you can prove that the normal distribution is a limiting case of the Poisson distribution with the parameter lambda going to infinity. I'm going to leave that too as an exercise. So, to use the distribution, we have to import it from scipy stats. It's going to ask us only for the mu, and it's going to ask us only for the size. This is the data Poisson here, and the plot is something like this. So, what's the difference here? The upper one is a continuous function. This is a discrete function. So, this is why we don't have like these lines here, stuff like that. This is also, this is actually not that kind of function. So, this is the way we define the Poisson distribution. The final distribution that is also a very important one is called the t distribution or the student's t distribution.

And this is a widely used distribution in hypothesis testing that we're going to see in the next video and plays a central role in the very popular t-test. A t distribution describes samples drawn from a full population that follows a normal distribution. So, this is very close also the normal distribution. The larger the sample of the t distribution the more t distribution resembles a normal distribution. And with this distribution, we have a parameter that we didn't use in the other ones, and it's called the degrees of freedom or df. And it can be defined as a number of values in the calculation that are free to vary without violating the result of the calculation. That's the definition of a degree of freedom. You'll see that a lot in the world of statistics. The formula for the probability, for the distribution is something like this. So, it looks kind of weird. I'm not going to lie to you, but if you check about this beta here, this is not a B, this is a beta, a capital beta, this is the beta function. And this mu is a positive integer that is called the shape parameter.

The formula for the beta function is this. So, it's an integral, if you don't know what is that, you can search for that in calculus, integral calculus, and you'll find what is an integral here, but it's the definition of the beta function, and this is called a special function, widely used in physics. So, to import it, we use import t and that's it. It's going to ask us for the degrees of freedom, again the mu, and the sigma. We do it here, and this is our data, but something is weird here. What's happening? This is not what we expected. Before, when we were creating this kind of stuff, with used the rvs, maybe that's the case. Because if you see here, I didn't put here dot rvs. So, let me do it again here with rvs. And no, we only have one point, because we're freezing in just one point. We don't want just one point. We want a whole distribution. So, to do that, we need to do something different for this distribution, and we have to create a linear space. A linear space, you can think of it as like a continuous base of numbers that can define a line maybe, and in here, we're creating one from 0.0001 to 0.9999, and we're creating 10,000 points of this. We do this, so again, the process is we create random variate from t.

We create a linear space, and then we create the pdf. So, you can search for all of these terms in the documentation of the library. It's very simple. When we do that, now we have an array. That's what we expected. And so, if we do a distplot, and remember we were doing distplots all the time before, if you do this, what you get is something like this. This is not actually a right plot for that. I mean this is not a wrong plot, but this is not the distribution plot you will expect from this type of distribution. And the thing here is that we don't want this distribution plot. We want to have a line plot. Because we created data in a different way here. We created a linear space, and so it's not the same as we did before. So, for us to be able to plot it correctly, we need to use a line plot. And as you can see here, this is very similar to just a normal distribution, and we will cover more about the t distribution in the next video. See you very soon.

****

Video 5: Confidence Intervals: Part 1 (12:53)

Welcome. Today we'll be talking about confidence intervals. From an outline perspective, we'll go through an introduction of terms. We'll talk about what confidence intervals are. I'll show you a little bit of Python code, and then we'll wrap up. So, from an introduction perspective, what is a confidence interval? So, a confidence interval is an estimate that's computed from the observed data statistics.

Meaning, if we have a sample set and we calculate statistics of that sample set, we can actually determine to what degree of confidence can we be sure that a particular value would fall within a confidence interval. So, in the vernacular, if we're 99% certain that a particular point or value for example, the mean, would fall within a confidence level, then we would say that particular assertion that we want to make with a high degree of confidence is true. So, in order to be able to do that, let's just take a look at a graph of what this would look like from a normal distribution perspective. So, the confidence interval here in this particular case, we want to be able to determine 95% confidence, just as an example, we would then be able to determine what are the particular values that should fall within that normal distribution which would allow us to be able to say with confidence the value that we're looking for would fall within that particular range. In order to do that, we would need to calculate what's called the z value.

And the z value is the value which indicates how far from the mean we are along a normalized distribution. So, in order to be able to do that, let me cover a few more terms and then we'll talk about the equations which allow us to be able to calculate those values for us of the confidence interval. So, the first one is what's called a z score. And as I just mentioned, the z score is typically referred to as the mean of a distribution, which is -- which is a standard or normal distribution.

So, in this particular case, if we wanted to determine something like temperature or whether it's your own temperature, you would then want to be able to follow what's called a -- the z score and then determine whether or not your temperature, say, falls within a particular range with a certain degree of confidence. Now, why would I use a normal distribution? The normal distribution is used typically when your sample sizes are large. And so, if you -- we'll talk a little bit about Central Limit Theorem, but basically, the Central Limit Theorem tells us that for any large dataset in real world phenomenon, most datasets would tend to exemplify a normal distribution.

And so, in that particular case, we would use the z values or the normal distribution to allow us to be able to determine what's our confidence that a particular value would fall within a range with a degree of confidence. So, that's fundamentally what we're going to do. If the sample size, though, is less than 30, or a small sample size, then we would use what's called the t-test. So, the t distribution allows us to be able to do something similar, but it's not quite as standard as the normal distribution.

So, in that particular case, we would have to specify the degrees of freedom that we would use, which would be predicated on the sample size, which would allow us to be able to change the shape effectively of our distribution which would be based off of the number of samples that we have available to us. So, let's just talk about the, at least the first case, let's talk about a standard distribution. So, in order to be able to calculate the confidence levels, the first thing that we want to do is be able to determine what our z score is. So, our z score, in the simple sense, is going to be defined by the sample mean minus the value that we're looking to determine divided by the -- or, determine whether it's in the conference range divided by the standard deviation.

So, if we're looking to create an interval, and I'm not going to get into a lot of detail related to the algorithm, but you can go through the equations themselves, but I want to stop at the bottom. So, effectively, what we're going to do in order to determine our confidence interval, or confidence levels, is to be able to use our dataset and then use our z values to be able to calculate both the lower bound and the upper bound, which is our interval of confidence. And so, the first thing you look at is the mean of the sample dataset, which is called xr, and then we would use the Z value, which we can look up in the z table or the z distribution table to be able to determine what the value is. In this particular case, we'll get to what alpha definition is as well, but the alpha, which you can see here in the equation, is really just the confidence level, or 1 minus the confidence level.

And because we are looking at both sides of the equation, let me just scroll back up here just a second, or sorry, both sides of the distribution. If we're looking at both sides of the distribution, then I want to ensure that I'm actually having the values here. So, in the particular case of 95% confidence level, you'll notice that I have 2.5% on the left side and 2.5% on the right side, meaning that the area under that curve here, in order to give me the 95% confidence level, is not 5% on either side. In order to be able to determine my confidence level or, in this particular case, either side of the distribution is the likelihood that it does not fall within the interval, it's -- we need to ensure -- make sure that we have it in order to get the upper and lower bound.

So, that's effectively what this equation does. So, here the plus and minus values are included, and then in order to complete the equation we take the standard deviation and then we need to include the factor related to the square root of the number of samples that have been collected or that are being analyzed. Now, in order to make this a lot easier, instead of actually going through the full calculation, there are Python functions which allow us to be able to do that. And so, we can use interval functions based on the type of distribution that we're using. We've been talking about the normal distribution, so we can actually go through and show you what that would look like. So, the first thing we want to do is just show you some libraries that we would use in order to be able to calculate that interval of the confidence level, or the confidence interval. And the first part is to include Scipy Stats Library, which has the important functions for us. And then we'll also include [inaudible] pi for other supporting functions.

So here, the first thing I want to do is determine the alpha value. I mentioned it just a bit ago. And in order to be able to determine what the alpha value is, because it's an important part of our calculations, we are just going to use the overall confidence interval or confidence level that we're setting, and then we're going to basically take 1 minus that value to determine the alpha value. So, that's relatively simple to understand what that is. So, let's just go through in Python and look at what this would look like in terms of being able to determine what the interval would look like. And I'm going to use the normal function. I'm going to calculate the interval, a 95% confidence level. And then I'm going to show you what this would look like if I use the percent point function as well to determine the upper and the lower values. Now, remember, in order to calculate upper and lower values, I need to effectively divide the values by 2. Again, going back up to the diagram here, upper and lower values would be these upper and lower confidence interval values, which would allow me to be able to then specify this 95% confidence range.

So, basically, I would have confidence that my value that I'm anticipating or looking at would fall within that particular range, excluding things that are outside the area under the curve beyond that particular value, where the z value is defined, and to the left, and here on the right-hand side, the particular value here, 1.96 to the right. So, that's exactly what we're going to be calculating. So, let's just take a quick look at that. And if we run the code we can see that if we calculated the confidence intervals here, we can see for 95% we can see the interval itself, which is minus point -- sorry, minus 1.95 and then positive 1.95.

And then, as I mentioned earlier, I can look at one or the other on the left or the right-hand side based on how I performed the calculation using the percent point function. And it would give me either the positive, or the right-hand side of the distribution, or the left-hand side of the distribution. And then in order for us to be able to calculate the interval with one calculation, which would make it a little bit easier, we can go through the process then of just calculate each one of the independent values.

So, here I'm setting my alpha. I'm going to set my upper value, which is my interval end value, and then then I can go through and calculate based off of my standard deviation, my mean function, which is going to be my mean values from my sample set, which I'm setting here to 135. And I picked the number of, let me just scroll up a little bit, I picked the number of values, which was included here as well, to be 56. This is not a real dataset, so I'm just setting the values in order to show you what this would look like. But also bear in mind, we talked a little bit about the -- whether or not to use a z or d -- or sorry, t distribution based on a number of samples. So, in this case, we want to make sure that we're using a normal distribution, because 56 items is greater than 30. So, we want to make sure that we're using that particular type of distribution.

So, let's also take a look then at what the conventional calculation would look like as we set the values. So, here I'm going to actually run through the equation itself that we were talking about earlier and I showed you, and then we're going to actually calculate the plus and minus values that would be the execution of the algorithm. So, for this particular dataset and the value ranges, which are based off of the mean, or the x bar value of 135, in addition to its standard deviation and according to the calculation, you can see that this particular interval range for 95% confidence interval is between 134.7 and 135.2. So, that's how you would calculated it if you wanted to do this by hand. But in most cases, we're not going to want to do this by hand.

So, what we want to do is use Python functions, which would allow us to be able to do this much more simply, by using what's called an interval function. And so, the interval function we would set the important values, which were parameters to the function, which would allow us to be able to effectively calculate the same set of values, but by using the interval function itself. And so, here you can see it gives me the exact same set of values, which are my, on the low side, 134.7, on the high side 135.2. So, what did we talk about today? So, we talked about the definition of confidence intervals, what they were, some of the practical applications of confidence intervals, and then how to calculate them by using Python functions, or by directly writing the algorithm in Python to be able to calculate confidence intervals. In the next section we'll talk a little bit more about how to do this with an actual dataset. Thank you for your attention.

Video 6: Confidence Intervals: Part 2 (10:47)

Welcome. Today we'll be talking about Confidence Intervals. From an outline perspective, we'll give you an introduction. Then we'll talk about determining the confidence intervals with the Z and T-scores. And then we'll wrap up. So, if you remember from part one of this particular video, we talked about the definition of a confidence interval. And as we mentioned, you can use a normal distribution or a T-distribution, which are very similar. However, the T-distribution, which we'll get into in a little bit of detail soon, requires what's called a degrees of freedom, because it is not completely normal distribution so you need to specify the degrees of freedom.

So, when you choose the type of confidence interval function, you would use the sample size in order to be able to determine that. So, normally, for small sample sizes or around 30, we would use a T-distribution score and if your sample size is larger than 30, then you would use a normal distribution. So, the assumption there is that larger data sets would have more of a normal distribution. So, if we talk a little bit about the Z and T-score to give you some more practical information. So, here if you should use a Z-score if you know the population variance and if not, use the T-score. So, here it sounds a little bit like a slight contradiction. But it's not, actually. So, what this is really telling you is that in most cases you're not going to know the actual population variance.

Because in order to know the population variance, you would need to know all of the data in the population. So, sampling actually means that we don't actually know. So, in most cases if you're doing sampling, you probably want to use a T-score because you may not have an exactly normal distribution. So, this is what we're going to talk about a little bit and part of the strategy that we want to use as we're looking at how we calculate confidence intervals is really understand what do we know about our population based on the statistics that we have available to us. And then we could use either, or, either or the T-or Z-distribution. And then be able to compare the differences to see what feels or looks more logical in terms of our assessment or understanding of the data.

So, in order to do that, let's take a look at some stat packages. And the stat package is a Python package, which you want to include and it includes the T-distribution and also, if you remember, the Z-distribution or the normal distribution. So, we'll be using both of these. And I mentioned earlier the degrees of freedom. So, the degrees of freedom actually will need to be set and it is defined as the number of samples in your dataset minus 1. So, if you had 21 observations, as an example, the degrees of freedom would be set to 20. So, let's get going and write some code. So, the first thing we want to do is set up our environment. And as I mentioned earlier, we're using the stats library.

So, we're going to get that from scipy. And we also want to use numpy for some of the functions that are available to help us with our calculations. We're also going to use seaborn, which is going to allow us to be able to plot our data. And so, we can set up some styles and plot sizes just to make things simpler when we want to plot our data. And seaborn is actually a very simple package for allowing us to be able to do very quick plotting without necessarily having to do a lot of code, which is why I like it. And the last part is we need to include our dataset. And so, in this particular case we're going to use a dataset, which includes body temperature, in addition to body temperature, it includes heartrate.

So, let me run this and the next thing we're going to do is just take a quick look at the data itself. And so, if we look at the top of the data, the head function, you can see that -- we can see the temperature, which is the temperature of the sample for the individual, the sex of the individual, and we're lonely looking at the top five rows. So, we're not seeing the female population. Zero, our male population, 1 is female. But we're only looking at the top five. But, trust me, there's 65 female samples in here. We'll get to that in a second. And then BPM is the heartrate. So, the heartrate actually represents the sample heartrate for that individual when that sample was taken.

So, the next thing we want to do is because in this case I want to use just the female population for our analysis, and so what I want to do is basically just select from the data frame where the condition, sex equal to 1 is true, and then I'm only using heartrate or BPM, and then I want to take a look at the data. So, let's just take a look at the data. So, let's just take a quick look at this. So, here I've pulled out all of the data from my data frame where the sex was equal to 1. And then I'm looking at, again, just a shortened version of the data but you can see that there are 65 items in this list. And all that's being printed right now is just a subset of that data.

But there are 65 items. So, what we want to do next is just use a -- I talked about seaborn earlier, I want to be able to take a look at what this would look like if I were to graph this data. And so, very simply now, what I can do is just use this plot from seaborn and then my pass, my data frame or set of data and just take a quick look at what the distribution of that sample data would look like. So, pretty simple but also very powerful because it allows you to be able to get a visual perspective on what the data may look like. So, we already know, as I mentioned earlier, that there were 65 samples in the dataset.

So, what I want to do next is to start to calculate some of my base variables that I'm going to need in order to be able to calculate the confidence interval and, ultimately, where I'm going with this is I want to get to calling the stat functions that we talked about for both normal distribution and the T-distribution, which are norm and T-functions. We'll get to that in a second. But in order to do that, I first need to calculate the end variable, which is the number of samples, which is really just the length of the items in the data frame, my sample mean, and then my standard deviation.

So, we can just take a quick look at this and you can see my mean and standard deviation for the dataset, which seems to make sense. Somewhere around 74 is approximately the mean with a standard deviation of 8, which feels like a relatively large standard deviation. But that's okay. We didn't control the data. We just have to look at the data. So, the next thing we want to do is we want to calculate the confidence interval using the Z-score, the normal distribution. So, in this one, here I want to be able to use the scipy stats function, which allows me to be able to call the Z-interval, the normal interval and then take a look at what the actual confidence interval would be if I'm using an alpha value of .05, which means 95% confidence interval.

So, then you can see I passed in X bar and then I'm passing in my scale, which is my standard deviation divided by the square root of the number of samples. So, take a quick look at this. And so, the interval here is between 74.09 and 74.21 beats per minute, which is a relatively tight -- relatively tight sample size or confidence interval size for the data. But that's okay because we're asking for 95% confidence. So, the next thing I want to do, so now this was based on -- this particular one was based off of the normal distributions. Now, I want to take a look at the T-distributions.

So, in this case, now, I want to make sure that I'm setting the degrees of freedom. So, the degrees of freedom will allow me to be able to then say based on the sample, the number of samples that I have, I can control the shape of the distribution. And that's an important characteristic. So, it's not specifically a normal, even though it's symmetric distribution, it's not specifically normal because the degrees of freedom allow you to be able to change it based on the observed number of samples. So, let's take a quick look at this first. And you can see here -- sorry. Take a quick look at that. You can see our sample mean, the number of observations which was 65, which we knew.

The standard deviation is the same that we calculated before with the Z-distribution and I'll talk about why that is in a minute. And then we could see, we used the confidence interval here which was 74.153 and then plus or minus .06. So, this is the symmetry that we're providing. But a more convenient way to look at it would be to use the actual interval, itself. So, let's just take a look at how we could use the interval and make it a little bit simpler to view and this way we could compare it directly to the normal function. So, let's take a quick look at that.

so, the one thing that you'll notice here, and if I scroll back up just a little bit, I'll get them both on the screen. You'll notice that both distributions are showing the, well, almost exact same data. Not 100%. But at least for the first two, first three places, looks like in both charts, is pretty close. So, the distributions are almost exactly the same. The intervals are almost exactly the same. Why is that? That's because our sample sizes are relatively large. So, we tend to be getting almost the exact same phenomenon from either one of these distributions. So, this is just an example of where the dataset is normal enough that it allows us to be able to use either one of the functions.

So, that's the kind of thing that you may actually see in the real world. You may end up having to compare, sort of, the T-distribution against the Z-distribution, or the normal distribution, in order to determine which is better reflective of the actual confidence intervals that you're looking to achieve. So, in conclusion, we talked about determining the confidence interval with the Z-score and the T-score. And we talked about the methods that you can use within Python, the stats package to be able to accomplish that. Thank you, so much, for your attention.

Video 7: Hypothesis Testing: Part 1 (02:08)

Welcome. Today, we'll be talking about hypothesis testing from an outline perspective. We'll go through hypothesis testing using Python, p values, and a number of other definitions that will be helpful as we discuss the topic, some Python examples, and then we'll wrap up. So, in general, the purpose of statistics is to test hypotheses which we all know. And so, here we want to go through that in a little bit more detail. So, what is a hypothesis? So, this is where we want to define two mutually exclusive statements, and then we want to be able to evaluate which one is true.

So, that's what we'll be setting up as we go. The statement that's favored or the one that you expect to occur is called the null hypothesis, and then the alternative hypothesis which is the antithesis is the opposite. And then, when we describe it, we'll relate it back to some things that you've seen before based on the confidence intervals, and then I will show you how it relates to the confidence interval as well.

So, let's just take a quick example. So, let's say we were talking about the effectiveness of a particular drug. So, in the particular case, we won't have all the data for all of the population that would want to be able to test this against, so we will use a sample of a population which would be, let's say, it would be our trial tests. So, we want to understand is whether or not the drug was effective or not.

So, our null hypothesis would be that the drug was effective, and then we want to measure against a sample dataset whether or not that is the case. So, that's what we would be testing. So, that's a good practical use case of how we would do it. So, the nomenclature and the thing that you'll see in terms of representation is null hypothesis represented by H subscript 0 and then the alternative hypothesis represented by H subscript A.

Video 8: Hypothesis Testing: Part 2 (13:51)

So, now that we've talked about the definitions of the null hypothesis and the alternative hypothesis, I want to talk about p values. So, here a p value represents the probability of generating observed data that's favorable to the alternative hypothesis. And so, the reason for this is we're looking for the cases where we would reject the null hypothesis. So, that primarily is what we're after. So, what I wanted to show you, let's take a look at this diagram here. So, if you remember from the conversations on confidence intervals, we were looking at confidence intervals representing the likelihood that a particular value would fall within the range of confidence along the distribution.

So, for example, if we were looking for, say, a value that would be within a confidence level of 95%, we would want 95% confidence that our value was within that range. So, now we're looking for something a little bit different because we're talking about hypothesis testing. So, now what we're really looking for is we're looking for the value that would lie outside of that range. And so, we want to understand, and that's called the p value. We want to understand what's the likelihood that we would reject the null hypothesis? So, let's just jump down to the detailed first. So, in the case of the two tailed, null hypothesis, don't forget, is set to equality or can be set to equality. So, if it was, if I was looking for, say, value a equal to value b for the null hypothesis, then the alternative hypothesis, if you remember, because it's mutually exclusive, then would be the antithesis of that which would be not equal.

So, on the bottom, it's called a two tailed test because now I need to look for the likelihood that the inequality would exist on either the right hand side or the left hand side of my distribution. And so, that's the, that's why it's called a two tailed test. In the other cases, the alternative hypothesis is validating for less than or equal to and then depending on which one for the alternative hypothesis, you would use either the right or the left tailed p value.

So, let's just go back up here and the actual calculation of the p value is so we'll go through how to calculate it, but once I've calculated that p value or that area under that part of the curve for the distribution based off of the samples that I have, then I want to compare that to another value which is called the alpha value. And so, this, if the p value is less than or equal to the alpha value, then it would be strong evidence against the null hypothesis, meaning that I would reject the null hypothesis. And so, the alpha value in this particular case is going to be based off of the confidence interval, but in this case, it's going to be one minus the confidence interval, and we'll talk about the alpha value in a little bit.

But, because we've already talked through this, the other thing that I want to mention just quickly is if the p value is greater than or equal to the alpha value, then we have strong evidence for the null hypothesis, meaning we would accept it. And, I'll talk about alpha value, what that is, but it's really just a simple concept of if you look at the degree to which your, you want confidence in your results, then you would set the alpha value to be equal to one minus that confidence level that you're looking for. And so, that's called a significance level. But, let me explain the test statistic first, and then we'll get into alpha value in a little bit more detail.

So, the test statistic is what we're going to use in the hypothesis test to decide whether or not to support or reject the null hypothesis. So, this gives us a value that we can use for comparison purposes. And, we'll do two different forms of this. So, the first one is going to be either a Z-score or a t score, and if you remember from prior discussions, it's based off of either the normal distribution which is our Z-score or a t distribution which is our t score. And, we'll go through the calculations of this as well. So, getting back to the alpha value here just quickly.

So, we talked about it being the probability of rejecting the null hypothesis when the null hypothesis is true. So, what does that mean? It effectively means that we have confidence that in this particular case, that if the null hypothesis is rejected, then we would say that we have a really small degree of error that that is true. So, that's really what we're shooting for. So, in particular case, let me give you a good example of this. So, if I'm looking for something like airplane engine failures, I want to ensure that if that value is really small so that it would indicate that in support of the null hypothesis would basically mean if the, say the engine will perform, I want to ensure that I have 99% confidence in that occurring, meaning that there's only a 1% chance that I could be wrong, meaning that it's a false positive.

So, you can understand why this could be very important or why you would want to make sure that you get the values right and in terms of being able to determine whether to accept or reject the null hypothesis. So, what types of tests do we run? I talked about this a little bit earlier, but one tailed or two tailed, this has to do primarily with the alternative hypothesis. So, if you're looking for equality, then in equality that is in the null hypothesis, then you're going to use a two sided test because the alternative hypothesis will be, will check for inequality. And then, likewise for greater than or equal to. So, it's pretty easy to determine which type of test we'd be running. Now, we talked about some of the things that you would need in order to be able to calculate the hypothesis tests and to validate them. So, we talked about confidence intervals.

We talked about the alpha values, and then we talked about p values. But, mostly in what we're talking through here will be the calculation for p value, and you'll see how simple it is. But, conceptually, you still need to understand those other characteristics in order to make sure that we can run this properly. So, let me just jump down to the standard calculation. So, I want to go through an example here, and my example is going to be for human temperatures. And so, I have a dataset which will allow us to be able to look at some data which is coming from a male and female population. And, what we want to look for is does my sample set reflect or represent a null hypothesis meaning that does my average temperature in my sample set, is that going to be equal to 98.6 which is usually the accepted temperature for a human.

So, because of this, and I'm getting back to the explanation I was talking about before, we want to run, because we're checking for equality, we can run a two sided test. So, that's what we'll be doing. So, the first thing I want to do is set up my environment in Python, and so in order to do this, first I want to set it up or I want to import pandas because we'll use pandas to be able to process our data. NumPy for some of our scientific calculations, and then the important one here is going to be coming from the SciPy stats package which is the normal function. The normal function will allow us to do z tests, and I'll also include SciPy. But, let's just import the data first. Then, we can take a quick look at the data. We'll just look at the very top of the data.

And so, here, I can see just the temperatures for the first five records, and so it represents the, in this case, this x indicator is zero or one. Zero is the male population. One is the female population. In our particular example, we'll just be using the male population. So, the first thing that I want to do here is as we look at just the male dataset, I can use a pandas function to be able to select the sex and then just set it equal to zero. And then, I can use these convenience functions to calculate the mean and the standard deviation which we're going to require in order to be able to calculate our test statistic.

And so, here, I have a 98.1 is the average or the mean, and then I have a standard deviation of .69 So, from there, what we can do is we can take a quick look at this and say our mean's off by about half a degree, and so is that sufficient to be able to reject the null hypothesis, which is fundamentally what we're looking for. So, the, what we're after here is we want to assume that this would be the thing that we're checking for is the sample dataset, the mean of that sample dataset equal to 98.6? And so, that's what we're checking for. So, we're going to use a normal distribution. The first thing that I want to check for, then, is I want to calculate my Z-score or my z statistic.

So, that test statistic would be calculated by performing the Z-score calculation which is basically just going to be the difference in the observed mean minus the expected mean, which is 98.6, and then divided by the standard error. So, if I run through that quickly, you can see that it's minus 5.7 So, from there, now what I can do, and this gets to the power of the p value. So, now, what I'm really after is I want to calculate the value under the, that point in the curve which would represent the in support or rejection of the null hypothesis. And so, in this case, then, I calculate my p value, and I'm going to use a probability density function using my test statistic, and then in this particular case, I'm multiplying by two because it's two sided probability function.

So, now, if I check the value, you can see that my p value is going to be equal to a really small number. So, what this means, then, is that if I compare that to my alpha values, so the alpha value was set in this particular case to .05 which was 5%. And so, here, you can see that at a 95% confidence level, the p value is very small. And so, that small value would indicate that we would reject the null hypothesis, meaning that it's not within our confidence level so that we would say, then, that the average of 98.1 would not be an accepted hypothesis for the mean or the value of that particular data against the sample set. So, the next thing we would do is then check the same thing using test statistic. Now, we don't have to use both of these, as I was talking about in prior conversations. The real difference normally comes in as a result of whether or not we know the standard deviation and whether or not we have a large sample size.

So, usually the t score is related to a small dataset. So, we would use a t distribution for a smaller dataset. And then, in that particular case, we need to make sure that we set the degrees of freedom. Degrees of freedom would allow us to then change the shape. Usually, it would make the tails on the outside of the graph a little bit flatter, but in this particular case, we're going to run through the example just to show what it would look like if we used the test statistic which is a t score. However, we do have 65 samples which is still a little bit large, then, for the t distribution. But, that's okay. We're just going to run through it and see what it looks like.

So, here, we're going to go through, include our stats function which we already had actually. But, in a similar way, we want to calculate the mean and then we want to calculate our standard deviation. And then, we can perform the calculation of the test statistic. So, that'll be the next thing that we calculate. So, first, we'll calculate the test statistic, and then we'll calculate the p value. Now, notice that when I run this, hang on one second. Once I run it for the two sided, you can see once this is run that what'll end up happening is in this particular case, the test statistic is similar in nature, not exactly the same because we're using a slightly different distribution, and then the p value, notice, is also small. So, this is what we would expect.

So, we would expect, again, in this case, a rejection of the null hypothesis, but the values are just a little bit different, primarily because of the fact that we're actually using a different distribution type. So, just in wrapping up, we went through the definition of what hypothesis testing was and how to do it in Python. We talked about p values and a number of other concepts which are in support of allowing us to be able to perform hypothesis testing. Then, we went through two examples related to using t scores as a test statistic and Z-scores as a test statistic. And, thank you very much for your attention.

****

---END--