Introduction to Data Analytics

Prof. Nandan Sudarsanam and Prof. B. Ravindran

Department of Management Studies and

Department of Computer Science and Engineering

Indian Institute of Technology, Madras

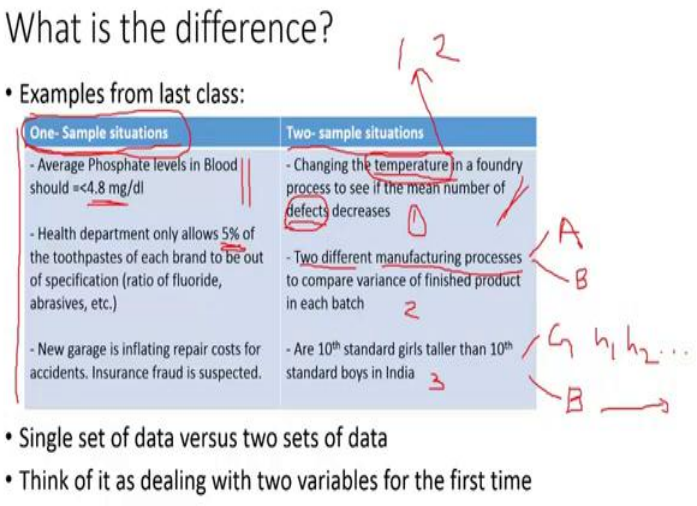
Module – 03

Lecture – 11

Inferential Statistics - Two Sample Tests

Hello and welcome to our next lecture in inferential statistics. So, today, we will continue our series of lectures and hypothesis tests; and specifically, we will build upon our previous lectures. So, just as a reminder, we motivated inferential statistics and the use of hypothesis tests two lectures ago and, in the previous lecture, we focused more on the single sample tests.

(Refer Slide Time: 00:37)



Now, if you remember from our previous classes – two classes back, when we provided examples, I spoke about the use of the one-sample situation as well as the two-sample situation. So, this table that you see in front of you is something that we discussed previously and we have gone over this. And, what you see on the left-hand side, the one sample situations – some examples; and, on the right-hand side, here are some examples of the two sample situations. In the last class, when we did some tests, we focused on the

one-sample case; and today, we are going to focus more on two-sample case. So, what is the big difference? So, again, just if you jog your memory, we spoke about various single sample tests.

So, let us go back to the single sample tests. In the one-sample test… I am using the word single in one sample interchangeably. In the one-sample situation, we were either testing – doing a hypothesis test on the population mean or we would do a hypothesis test on the proportion – a proportion – again associated with the mean proportion – associated with mean; and, or sometimes we would do a hypothesis test on the population standard deviation. But, in all these cases, whether it is a proportion that you are testing, whether it is the mean that you are testing or whether it is the standard deviation slash variance that you are testing; in all these cases, you always had a single dataset. Your sample came from a single distribution. And so, you had a single sample in your hand. A single sample does not mean a single data points; you had many data points, but the data points represented one particular distribution or one particular context. And, in all these cases, you would invariably compare this dataset to a single number that you had in mind.

So, you would test the hypothesis that, the average phosphate levels in blood were less than 4.8. Here the dataset that you had was the dataset associated with phosphate levels in blood for may be a person or a machine or a set of people; whatever the context is, there is still one dataset and that set was used and compared to a specific number. In this particular case, the specific number was 4.8; that number could be something else. In the case of the proportion test with the health department, it was 5 percent that you were testing against. And, in different cases, in the case of the garage, it was comparing it to the national average, which conceivably would have been a particular number. So, that is the third example. But, in all these cases, whether you are testing for mean, whether you are testing for proportion or whether you are testing for standard deviation, just keep in mind that, what makes it a single sample test is that, you do not have multiple sets of data; you have the single set of data.

And, you are always comparing that set of data to the set of data you sample. But, using that sample, you are saying something about the population. And, what you are saying about the population is essentially a comparison with a particular number that you have in your head. So, you are trying to see if it is less than 4.8. So, those are the single

sample cases. But, in the case of the two samples, you actually will get two sets of data. And, typically the way it happens is… For instance, in the first example, you see there in this table, you actually go change the temperature; so, you actually go mess with a particular variable and then you look at the number of defects. So, you have dataset one, which might be a set of 10 or 20 or 30 data points and the number of defects in different casts. And then, you went and changed the temperature and you get another set of 10 or 20 or 30 data points. Sometimes these numbers are not always equal – meaning – cannot have the same number of data points on either side, more often they are not; it is good that you do.

But, the core idea is that, you now have two sets of data that represent two separate distributions. Distribution 1 is for the number of defects that you would expect to see before you change the temperature. And, distribution 2 is the number of defects you would expect to see after you change the distribution, after you change the temperature. And, these two distributions correspond to their respective populations. And, what you are doing is you are taking a sample from both of these populations and near comparing these two samples and you are ultimately trying to make a statement about the mean of population 1 versus the mean of population 2. So, that is essentially the first example that you see here.

Now, the second example is again one where nothing was changed over time; but, let us say you just had two parallel different manufacturing processes. And, this is a case, where you want to compare the variance; you do not really care about the mean of the manufacturing processes. So, let us say these two processes are coming up with some finished product. You are more interested in the variability of this finished product. So, you do not care really what the mean is. But, again the idea said you will have manufacturing process A and there will be some distribution associated with it. And, there will be a dataset associated with that sample and you will have manufacturing process B and you will have another dataset associated with that sample.

And, even the last example which is tenth standard girls taller than tenth standard boys is also fairly is just a straightforward extension of what we discussed, where you have two different datasets and you are comparing the two datasets. So, keep in mind that you do not have to have a number in your mind. And, I will talk about the odd case where you do have a number in your mind; but, the one big difference is with two sample situations,

you are ultimately having two separate datasets. So, instead of thinking of the words as one sample and two sample, that sometimes gets confusing. I just like to think of it as a single set of data versus two sets of data; that kind of emphasizes that there are many data points. The other way of thinking about it, which sometimes helps me is just think of it as for the first time, the two-sample test, you are dealing with two variables; you are always going to have the variable that you are measuring. So, in the case of the…

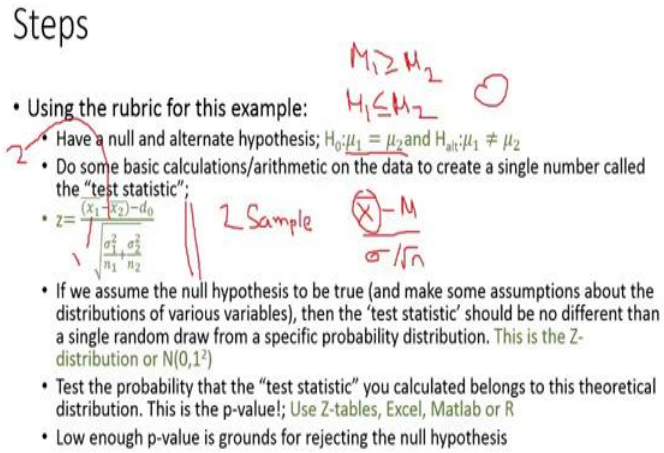
Let us take each of these examples. In the first case, the variables – I am going to call them 1, 2 and 3. So, let us do that. So, this is 1, this is example 2; and, this is example 3. You do not have to circle it. So, in example 1, the variable that you are measuring is the number of defects. So, that is your output variable. So, in some sense, you can say you have got one variable, which is the number of the variable as a number of defects. But, in addition to that, you have another variable, which is the temperature. And, the temperature takes on only two values. And, that is why it is a two-sample test. The temperature takes on the value that it was before you changed it and the value – it is after you changed it. So, for the first time, you are seeing two variables.

When you go back to the single sample situations; so, if you take the average phosphate levels example, the phosphate levels in blood was your response; so, essentially your output variable. And, that was the only variable that was involved; there was nothing else that you were changing. For the first time in the two-sample case, you will have two variables, which is the number of defects, which is your output, your response variable, your standard variable that you always see. And then, there is another thing that is changing, which is the temperature. So, the temperature before… And, the temperature can take on only two states: 1 and 2. And, those are the two datasets before and after; same thing with the manufacturing process to compare variance of the finished products; so, the variance of the finished product, the variance of a certain number. So, it might be the dimensions of the product of which you are interested in the variances of. So, you are interested in the variance of the dimensions of a certain product.

Let us say that was the diameter of a finished product. So, the diameter itself is the output variable and you are concerned about the variance of that. So, that is your output variable. But, the manufacturing process is your other variable with two difference states. So, there are two different manufacturing processes. So, if you call it manufacturing process A and then manufacturing process B; then, manufacturing process becomes your

second variable. And, this variable can take on only two states: A and B; you can call it 1 and 2 or whatever you want. Again out here in the third example for instance, what is your output variable? Quite straightforwardly height. So, you are measuring height in both cases; and that is your output variable in the third case. And, the second variable of interest; and, you can think of it as the input variable is gender. So, girls verses boys. So, you have datasets for girls and then you have a dataset for boys. Tenth standard is not a variable, because it is consistent across both. So, it is just a detail in some sense. But, what you will be measuring is height 1, height 2 and so on dot dot dot; and, same thing out here also. So, it is like you have… One way to think of the difference is obviously that you are dealing with the single dataset with the single sample case and you are always comparing that to a number. In a two-sample case, you are dealing with two datasets. And, another way to think of it is… For the first time, in the two-sample case, you will actually be encountering two variables. One – very clear output variable or the response variable that is the variable of interest, that is, what you are comparing. And then, the input variable; essentially, what are the two classes that you have created in your system that you are comparing this response variable across.

(Refer Slide Time: 10:42)



But, otherwise, you will find that the core steps; this rubric that we created last time in terms of steps for hypothesis test statistics, which is that you still need to have a null and alternate hypothesis. In this case, the null and alternate hypothesis looks a little different. If you remember, in your previous example, you would have a null hypothesis such as a

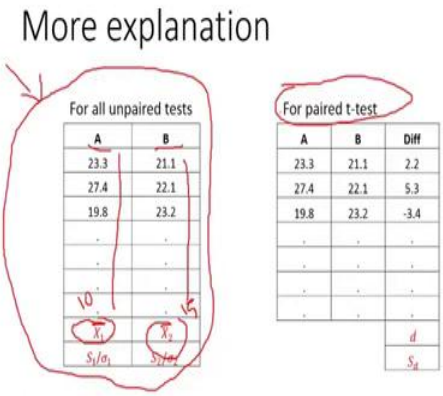
mu 1 is equal to 4.8 or mu 1 is less than or equal to 4.8. Here you will typically have a hypothesis like mu 1 is equal to mu 2; you can have different types of null and alternate hypothesis. If you remember in the single sample class, we spoke about how you can have a single tail tests and two-tail tests, where typically something going back to your single sample class thinks like mu 1 is equal to 4.8 would lead to two-tail test. And, we explained how that works. So, the same thing applies here. When you say mu 1 equals to mu 2, that is a two-tail test; whereas, when you say mu 1 is less than or equal to 4.8 or mu 1 is greater than or equal to 4.8, that would be one single-tail test. Similarly, out here you can have something which says mu 1 is less than or equal to mu 2; that can be null hypothesis. And correspondingly, the alternate hypothesis will also change and you can also have something that says mu 1 is greater than or equal to mu 2. So, all of these are possible.

But, accordingly, just remember – you will calculate the same test statistics. This formula will not change. But, accordingly, where you are, which part of the distribution you are interested in will change. And, if you still have some doubts about how that works, I strongly suggest that you go back to the lecture on single sample test and see how, which part of the distribution gets covered. And, we will also try to support that with some problems in your science. But, the important thing for you to remember with the two sample case is that, the same core concept of single-tail and two-tail tests hold. There will be a small exception to this and we will quickly cover that. It is not really a exception, but it is more of an addition and we will cover that towards end of this lecture in terms of creating more complex a hypothesis. But, for now, just seals as a simple extension of the single sample tests; but, you have a mu 1 and a mu 2.

The second step still hold, which is that you will do some basic calculations or arithmetic on data to create a single number called the test statistics. Last time we solved different formula for the single sample z-test. Here, you have given your formula for the two-sample z-test – two-sample z-test. And, the core idea here is for instance that… So, let us just go through this formula to get you some idea. The idea as such you have an x 1 bar and an x 2 bar. So, it is different from your old formula of x bar minus mu divided by sigma by square root of n. Now, this is the formula that you would have seen for the single sample z-test, because you had a single dataset and you have calculated an x bar

from that; and, mu was a number you had in mind. Here you have two datasets: 1 and 2. So, for both these datasets, you are going to calculate x 1 bar and x 2 bar.

(Refer Slide Time: 14:12)



So, just to give you some idea, I mean I have represented as a table. Just for your convenience, just focus on this table. We are going to use this table for some other purpose. So, on this table, with the arrow, you have two sets of data. So, one set of data corresponds A; one set of data corresponds to setting B. You can think of it as the boys and the girls or you can think of it as manufacturing process A versus B or whatever it is. And, this number out here in the bottom, which is x 1 bar and x 2 bar – essentially, there needs to be a small dash above it, if it is not clear. So, there is x 1 bar and x 2 bar. And, these correspond to the average. So, x 1 bar corresponds to the average of these numbers; x 2 bar corresponds to the average of these numbers. So, you can think of it as x a bar and x b bar, but you can also… x 1 bar and x 2 bar are also just fine; it is just convenience. So, this x 1 bar and x 2 bar… Let me clean this page up.

x 1 bar and x 2 bar is essentially what gets plugged into these two formulas. And, I am going to come to d naught in a second; but, similar to the z-test of the single sample case, you will need to know the standard deviation of the populations associated with the two means. So, x 1 bar and x 2 bar are essentially the sample means that you get from this distribution. But, you have this distribution associated with 1 and 2 separately with A and B separately. And so, if you are going to stick with the nomenclature of sample 1

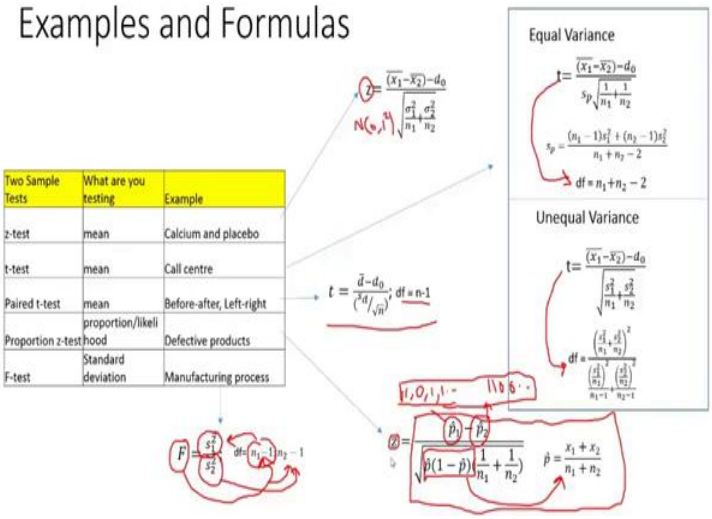
and sample 2, sample 1 has a mean – mu 1; sample 2 has a mean mu 2. We are testing the hypothesis that mu one is equal to mu 2; but, there is a standard deviation associated with sample 1 and a standard deviation associated with sample 2. That is sigma 1 and sigma 2. We need to know sigma 1 and sigma 2. So, those are numbers that need to be given to us. So, you need to be able to derive it from first principles, not from the data. So, you can plug those in; and, n 1 and n 2 would correspond to the number of the data points that is there. So, if these are 10 data points that are here; then, you would actually see. So, there are 10 data points in A; maybe there are 15 data points in B. So, you would plug in that 10 and 15 correspondingly and that would be n 1 and n 2. So, you plug all of that in.

And, I said I will come back d naught. And, this d naught is this small addition to the simple hypothesis that you can form here. So, if have a simple hypothesis like mu 1 equals to mu 2, then your d naught would be equal to 0. This is the idea where you say something like manufacturing process A; that mean of the manufacturing process A is equal to the mean of the manufacturing process B. And then, that is fairly straightforward. But, what if you said something a little bit more complex like the mean of manufacturing process A is 3 units higher than the mean of the manufacturing process B. In that case, your null hypothesis is really that mu 1. Suppose you were to say mean of manufacturing process A is equal to 3 units; mean of manufacturing process B plus 3 units. So, it is three units greater than mean of manufacturing process B. So, that becomes your null hypothesis; then, your null hypothesis becomes something like that, which is mu 2 plus 3 is equal mu 1; and then, d naught will take on a value. And then, d naught would take on the value 3. So, you can create more complex hypothesis. More often than not, d naught tends to be capital 0. You are more interested in questions like is A equal to B. But, sometimes if you are interested in questions like is A 3 units greater than B or is A equal to B plus 3 units just to keep it consistent as a two sample test. But, if you are more interested in saying is A greater than B plus 3 units. In those cases, again it will become a single sample test, but you would still need the user something like 3.

So, now, you do this and you calculate test statistics the same way as you did before. The formulas are little different. But, once you calculate it and you have a z number out here, the distribution you are talking about is the same distribution that you were dealing with in the single sample test, which is… So, if the null hypothesis is true, which is that mu

naught is equal mu 2; mu 1 is equal to mu 2. And, you make some assumptions; and then, the test statistics, which is what you have derived z should be no different than a single random number from a z distribution or a single normal distribution with mean 0 and standard deviation 1 square. So, that part of the logic is exactly the same. And, the way you test the probability using either the z tables from that is, these are some tables that you see in the back of statistics test books or you could use Excel or Matlab or R. And, we discussed some formulas in excel during the previous class. Those formulas are all identical; you are doing the exact same thing. And, the core idea is that, you use the software to calculate the probability, which is called a p value. And, if the p value is low enough, then you reject the null hypothesis; you reject this hypothesis that you created, which is that mu 1 equals mu 2. So, guys, like everything out here in this part is exactly the same as your single sample test. The only thing that change is this formula and your null and alternate hypothesis. Again the core concept of using two-tailed and one-tailed tests also is the same.

(Refer Slide Time: 20:13)



So, having described this, let us briefly go into the tests that are actually available. So, we have already discussed the z test. Another example that I can think of if you kind of learn by examples is… So, for instance, it is often thought that calcium could reduce – calcium supplements could reduce blood pressure. So, you could conceive of a simple test, where you give calcium to some people, let us say 20 people; and then, for others, you give placebo. So, people do not… These are just sugar coated pills where they think

they are taking calcium. And, you can compare… So, let us say you gave it a 20 in 20 people; you can compare the blood pressures of these two sets of people. So, you will have one set, which is a calcium set and then you have one set, which is the placebo set; and then, you can test the hypothesis that the mu of calcium… So, mu of calcium is equal to the mu of the placebo. Or, in this case, because there is the hypothesis goes out and says that calcium should create lower blood pressure, you can test the hypothesis that mu of calcium is less than or equal to mu of the placebo.

So, just another random example for you on that; you can alternatively use the t-test. And, the reason for using the t-test – again the goal of the z and the t test are always the same with the same difference as in the case of the single sample tests. The difference being that, in the t-test, you do not know the standard deviation already. In the z-test, you know sigma 1 and sigma 2. This is given to you in the t-test. This is not given to you. And so, you have to calculate s 1 and s 2 in order to figure out. Now, if you remember, in the single sample case, we said there is an exception to this rule. So, we said the basic rule is if you know standard deviation, use z; and, if you do not know standard deviation, use t. And, in most cases in life, you are not going to know the standard deviation. So, the t tends to be popular for that reason; I mean think about it right; it goes back to saying you are actually testing the hypothesis associated with means because you do not know mu 1 and mu 2. If you knew mu 1 and mu 2, you would not be doing any of this if…

Now, imagine a world, where you do not know mu 1 and mu 2; which means you do not know the population means of the two distributions. But, someone comes and tells you what the population standard deviations are; kind of rare to find. So, more often than not, the t gets used; but, we also spoke about this exception, where if your dataset is really large; and, large was defined by greater than a dataset size of 30, then you can actually compute the standard deviation and then the t-distribution approximates to the z-distribution. So, you can still use the z-test. If some of you find that confusing, you can keep it simple; you know the standard deviation, use the z; you do not know the standard deviation, use the t; you cannot go wrong with that.

Now, there is one small complication. And, I know these formulas can sometimes be a little scary. But, we are going to step through each of them. In the t-test, there are two versions; there are actually multiple versions. And, there is another version called the

paired t-test; and, we are going to come to that. But, we are not talking about that now. For now, I am talking about a straightforward t-test just the way you did the z-test for the same purpose of defining the difference in mean. But, in this particular case, in the t-test, you need to figure out whether the standard deviations or the variances are equal or not on principle. If you can say the standard deviation should be equal, you would use this formula, which is the equal variance formula; if you do not and you believe the standard deviation should not be equal, then you would use the unequal variance formula. So, the quick idea is that, the idea of x 1 bar, x 2 bar and d naught is the same as in the case of the z-test.

And, you use a single formula for standard deviation because its equal variance; and, that you get from computing s 1 square and s 2 square, which is nothing but the variance that is computed from the actual dataset. So, here is the dataset. And, this s 1 and s 2 are actually the standard deviations that are computed from this data. By the way, just for your reference, the sigma 1 and sigma 2 what you see here are not computed from that data; I have just put them under A and B to tell you where they belong; but, sigma 1 and sigma 2 are given to you as s 1 and s 2 are the standard deviation of… Like for instance, s 1 is the standard deviation of the this dataset; hope that make sense. So, let us also clean up this table grid. So, when you have equal variances, you can compute something called… And, this is called a pooled variance and that is what you will plug into this formula. And, we have also covered the concept of degrees of freedom. And, the degrees of freedom are nothing in this case, but the idea of the total number of data points. So, using the total number of data points to figure that out.

And, you will see that you can use a similar formula for the unequal variance case, where you will not have a concept of pooled variance and you will have a separate s 1 and s 2 square and the formula for degrees of freedom also differ. Again if you have any questions on degrees of freedom, feel free to refer to the previous lecture; that should give you a similar insight. You will also notice that again in both those cases; this is extra term d naught; sometimes some texts will not even include this formula; but, the idea is that, if your null hypothesis is quite straightforward, which is x 1 equal to x 2, or is x 1 less than or equal to x 2 or is x 1 greater than or equal to x 2; then, d naught is equal to 0. But, if there is an offset such as is x 1 equal to x 2 plus d naught. Then, you will need d naught. Or, is x 1 greater than x 2 plus d naught.

We will now move on to the next form of t-test. And, this is called the paired t-test. It is also trying to test the same core concept, which is mu 1 equal to mu 2. But, it is doing so in a slightly different way. And, the idea behind the test is the following. In a typical test, you have two datasets. And, here I am referring to everything that you can see on your left-hand side. Dataset A and you have dataset B. And, if you are trying to see if mu A equals mu B; and, you are doing that using x 1 bar and x 2 bar and so on. Now, if there is some kind of logical pairing between the rows; so far, we have actually ignored any connection between this data point to this data point, because there need not be any connection. And in fact, the number of data points in B need not be equal to the number of data points in A; which is why we have two completely different terms n 1 and n 2 in all these formulas. But, if n 1 was equal to n 2 and there was a logical connection between each point, then you want to use something called the paired t-test.

What do we mean by a logical connection between the two points? We mean the following. Either that… Let us say you are doing this same test of the calcium and placebo. We discussed this test in the context of a z-test. Suppose you did not know the standard deviations, you could have very well used it as a t-test. But, let us say that you were doing a calcium and placebo; but, the way you were doing the test was that, for each person – for each individual person, you would give a calcium tablet; look at the change in blood pressure; and then, on a separate day, give the same person the placebo tablet; that means, for each person, there would be one recording in calcium, one recording in placebo; which could mean that, each row could signify the calcium for person x and the placebo for person x. So, there is actually a very clear logical pairing. So, the second row could be something quite simply calcium for person y and placebo for person y. So, in that sense, these two data points while…

Yes, A continues to mean calcium, B continues to signify placebo; but, this is specifically for… There is a logical connection that both of these are connected by this person y. So, lot of times, applying the same treatment on the same x essentially experimental unit could be that logical paring. And, when you have that logical paring, a great way to get more out of your test is your paired t-test. And, the idea here is that, instead of computing x A bar and x B bar, x or B; used to call it x 1 bar and x 2 bar, we instead take the differences of A and B. And, each difference is computed here. So, this is the difference between 23.1 minus 21.1. And, that is 2.2. Once like that you compute

all these differences; and then, you get a d bar, which is nothing but the average of these differences. And, you get an s d, which is nothing but the standard deviation of these differences. And, you go ahead and plug that into this formula. So, you have this d bar s d – square root of n is just the number of data points and this d naught is the same concept as the d naught here; where, if you are saying if A is A equal to B, then your d naught is 0; if A equal to B plus 3, then d naught is 3. But, outside of that, d naught is the same core concept; but, the idea here is that you are using d bar s d instead of x 1 bar, x 2 bar separately.

You can think of other common examples about a logical pairing. For instance, if a common example is – if we are interested in knowing – if a particular track creates more wear and tear on the left-hand side of the particular tyre; let us say you are on a formula 1 context; it is a sport and you are really captain of a particular track creates more wear on the left-hand of the tyre than the right-hand of the tyre; then, you might do a two-sample test and you might compare left tyres and measure their wear. So, wear just means how much they have eroded. So, let us say there was some logical way of measuring that. Then, you would measure the wear on some sample of 20 left tyres and the wear on 20 right side tyres. Now, if these 20 left-hand side tyres and 20 right-hand tyres came from completely different cars, you would have to stick to your standard t-test.

However, for each car, if there was one left tyre and one right tyre as a data point; and then, you had 10 cars or 20 cars. So, you had 20 left tyres and 20 right tyres; then, you can use the paired t-test, because for each particular left tyre, there is one corresponding right tyre. You could use that logical association to use the paired test. We then move on to the equivalent of a proportion test. We saw this in the single sample case; but, again it works the same way as your old proportion test; and, that you have two samples here and you want to see if the number of defects through process A is worst than process B. So, again you are doing this comparison not against a particular number; but, you are doing this comparison between two samples and… But, you are comparing the proportion that you see in both the samples. So, that is what you are doing in the proportion z-test.

The formula here is quite straightforward. The proportion associated with sample 1; the proportion associated with sample 2. Always remember that, in a proportion test, your data is essentially extreme of 1s and 0s. So, you are not getting the actual numbers; you

are getting either yes's and no's or males and females or whatever; it is that, you are measuring. But, you are getting this p 1 comes from a dataset, which is 1, 0, 1, 1 dot dot dot. And, that p 1 is nothing but the average. And, p 2 is similarly coming from another dataset 1, 1, 0, 0, dot dot dot. And, for the sake of variance, you would use this. This is kind of like the concept of pooled variance. And, that p hat in general comes from this, where x 1 and x 2 are the number of 1s that you see overall. And, n 1 and n 2 are the total number of data points that are there overall. So, that is a fairly straightforward extension of the proportion z-test.

And, the final we come to is the F-test. The F-test is used when you want to compare the standard deviation of dataset 1 with the standard deviation of dataset 2. You can conversely think of it as comparing the variance of dataset 1 with the variance of dataset 2 because that is essentially what you are doing. And, the idea is that, you take sample variance from dataset 1, sample variance from dataset 2. And, that should logically… This is very similar to the chi square test; but, that should logically give you something called the F-distribution. This is the first time we are saying that distribution called F-distribution. But, the one difference between the F-distribution and some of the distributions we have seen so far is that, the F-distribution has two parameters that define it and there are the two degrees of freedom. So, it is actually called numerator degrees of freedom and denominator degrees of freedom. And, numerator degrees of freedom is defined by n 1 minus 1, which is what is a number of data points on dataset 1, which is what goes on the numerator. And, the denominator degrees of freedom is n 2 minus 1, which corresponds to the s 2 square that you calculated, which is the sample variance of the dataset 2.

Just to again recap for you the concept degrees of freedom with a z-distribution, that concept does not exist; the z-distribution is nothing but a normal distribution with mean 0 and standard deviation 1 square. With the t-distribution, by definition, t-distributions have mean 0. But, to describe exactly the t-distribution, you are talking about you need to signify the degrees of freedom; and, that gets signified by these formulas for equal and unequal variances. And then, you have the special cases of the paired t-test; where, again the degrees of freedom is n minus 1. Again with the proportion test, because of the binomial approximation to the normal, you are only getting a z-distribution at the end of it. So, ultimately, you will be – you use this test statistic; but, once this test statistic is

computed, you are still dealing with a normal 0 – mean 0 standard deviation 1. And finally, for the F-test, which does a test of compares two variances to see if they are equal or not; you have to define it based on numerator degrees of freedom as well as denominator degrees of freedom.

I hope that was clear and that you have a good feel for the use two-sample test. Again there is a lot of software out there, where you can just plug in dataset 1 and dataset 2. So, for instance, you might be able to just completely dump the entire dataset in its native format and tell the software what test to do. But, using these test statistics you have a better understanding of what you are doing when you understand the formulas. And, once you compute the formulas and carry out your tests, then you can use a software to get the exact probabilities.

Thank you and look forward to seeing you in the next lecture .