

Applied Econometrics
Prof. Tutan Ahmed
Vinod Gupta School of Management
Indian Institute of Technology-Kharagpur

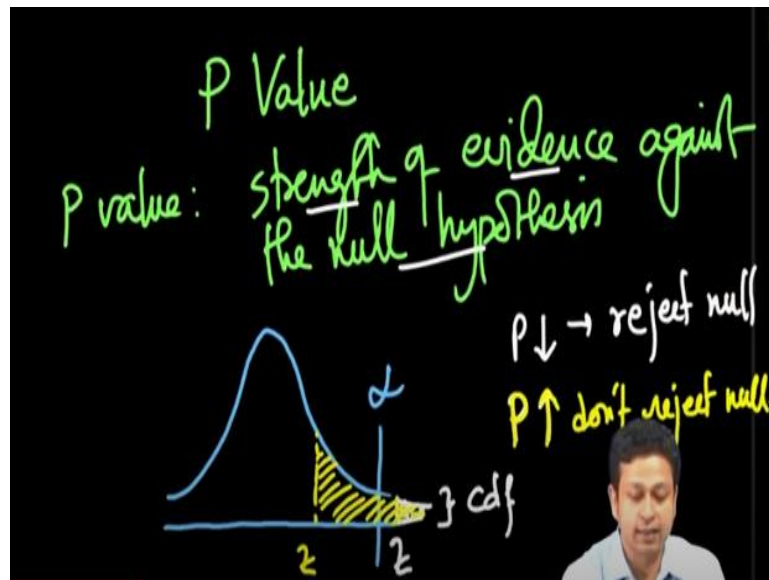
Lecture - 35
P Value

Hello and welcome back to the lecture on Applied Econometrics and we are talking about hypothesis testing. Now we have been talking about the decision rule for inferencing to do statistical inferences, and we said that we need to be familiar with some concepts to actually understand how we can actually do the inferencing. And we talked about several concepts like level of significance.

We talked about confidence level, confidence interval, P values, power and so forth. So in this lecture, we are going to sort of talk in detail about P value. And you will notice that whenever you talk about statistical inferencing, this concept P value comes you know pretty frequently and we often say that P value is so and so. So I decide to you know do so and so you know like I take so and so decision, we often do that is we often do it mechanically.

But there is a very interesting intuitive reasoning or logic how we see the P values and why it matters. And in this lecture, we are going to see the you know what the P value is as well as why we actually use something like a P value. And we will also do some sort of small hands-on to actually see how we can derive the P value. So let us start.

(Refer Slide Time: 01:42)



Now let us say, I have, I can define P value as we call it the strength of evidence against the null hypothesis. I repeat, P value is the strength of evidence against the null hypothesis. And we will see, the way we sort of use P value is let us say I have a distribution here. So very bad normal curve I have drawn. So let us say I have a distribution here. And for this distribution, let us say I have this level of significance alpha here and I have a P value.

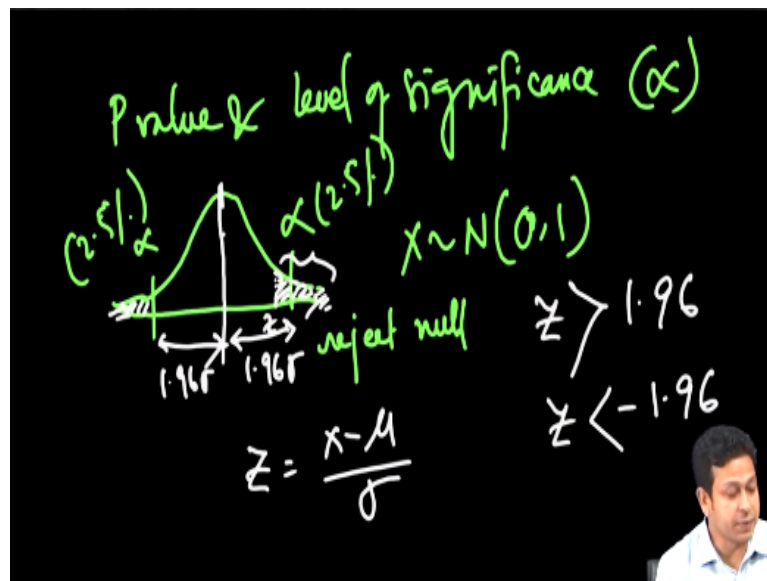
Let us say this is the P value here, I obtain the P value. And depending on my z, I will have my P value. And if I have a very small P value here, let us say this is the you know this is the CDF cumulative density function, we will just talk about it, for z. And if the P value is very small, then we say that we reject the null hypothesis, okay. So again going back to the wordings we have, strength of evidence against the null hypothesis.

So for a very low P value, low P value, or I will write down here, for a very low P value, we reject the null, okay. And for a high P value, let us say my P value is here, I will use a different color, let us say my P value is, my z is here. So the P value would be this much, the whole area under this curve. And then I will say I, so this is the, so I will use the yellow color here. If the P value is high, so I will say do not reject null.

That is how we sort of use the P value in taking the decision, right? So I just, this is the mechanical part of the P value, okay how we use it usually. But we need to

actually understand where from it has come. But before I go there, I will also explain this difference between P value and level of significance.

(Refer Slide Time: 04:11)



P value and say alpha or level of significance. So they are not the same. Sometimes people confuse the alpha as you have seen it is a predetermined thing. It is a predetermined thing. Alpha you already decided. So it is a 5%. It is a 1% alpha, whatever level we sort of want to you know whatever level of type I error we want to allow completely based upon the researchers judgment, the alpha could be chosen 5%, 1%, 10% whatever that is.

Whereas the P value that is obtained from the data, the researcher does not have any prior about the P value that he may get. It is just completely based on the estimation from the data. So as long as you know like so essentially it is this alpha. So as long as the P value is less than alpha, if it is a right-tailed test, or P value is greater than alpha, if it is a left-tailed test.

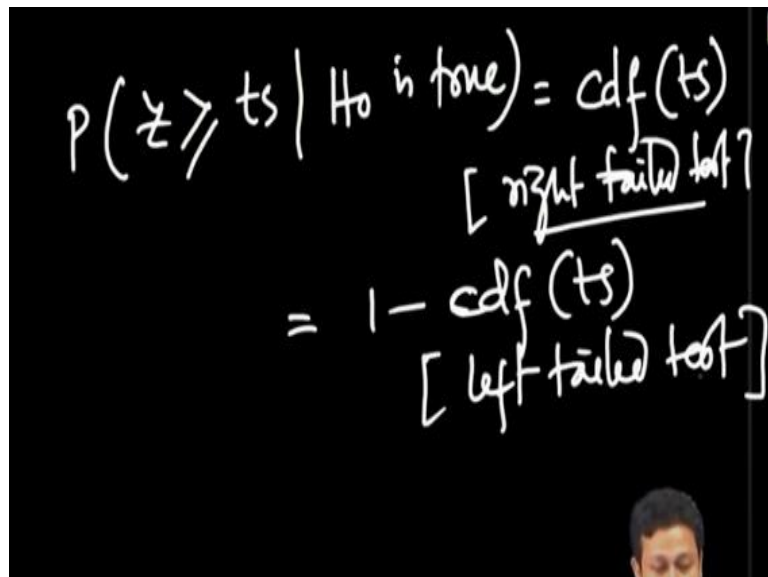
So we can basically say that we reject the null hypothesis. So we use both the P value and alpha to actually take a decision. So let us say if we have a 5% level of alpha, which means this, and for a two-tailed test, which means this for 5% level for a two-tailed test, the value of alpha is going to be 2.5%. And if let us say we have a standard normal distribution, so where x follows normal distributions 0, 1.

So what we will have? We reject the null, reject null. So basically if my z is either here or the z is either there right? So we will say reject null if z is greater than 1.96. So this is the value we have for this area, right? This is the standard deviation. The sigma is 1 and 1.96 sigma right, both side. So if z is greater than 1.96 sigma, so that means, if z is here we will basically reject. We also reject if z is less than minus of 1.96 right?

So essentially if z goes in this side. So in both the cases we reject the null hypothesis. Essentially that is how we use P value and level of significance to take a decision. Now we come to the y part.

Basically how we are actually coming to this understanding, you know how come we are using a tool like P value and how come we are coming to this decision if P is more than alpha or less than alpha we sort of take a decision if we take a call. So to do that, to understand that, we have to basically see the definition of a P value.

(Refer Slide Time: 07:16)


$$P(z \geq t_s | H_0 \text{ is true}) = \text{cdf}(t_s)$$

[right tailed test]

$$= 1 - \text{cdf}(t_s)$$

[left tailed test]

And P value is basically it measures the area under the curve for I will write down is the probability that the z statistic is, the value of the z statistic is more than the test statistic given that H_0 is true. So what I mean I will explain and that essentially would mean the CDF of the test statistic, okay. So in this particular case, this is what the right-tailed test.

What I mean, essentially if you see again, if we go back to the previous diagram, so if the value of the z statistic is more than the test statistic, okay. So if the value is for a given H_0 for a when the H_0 is true, I am basically whatever is the test statistic, whatever so if the test statistic is here, the value that I get, this cumulative density function if we just sum it up basically, that is the area under the curve.

So that will represent the P value. So here, my test statistic is z . So z is nothing but x minus μ by σ . So essentially whatever the value of z I get, so the area right to this is the P value, okay? And this is specifically that is why I have written that it is for the right-tailed test, okay and so the right-tailed test. So when I am assuming H_0 is true, okay?

So I am assuming it is a true sort of distribution, the sample is actually drawn from the true population. And then only I can, then I can actually sort of with this assumption, I estimate this area under this curve, right? And the area under the curve is basically the, it gives me the P value. Whereas for a left-tailed test, it will be one minus CDF for the test statistic, okay.

So it is quite obvious. For a left-tailed test what we will do, we will basically measure the area this side, alright? So basically, you have to understand that what is happening here is that again, the P value is actually measuring the extent from the actual true z we observe from the population or the sample, the z statistic that we get. And it is simply you know checking where the z statistic is lying.

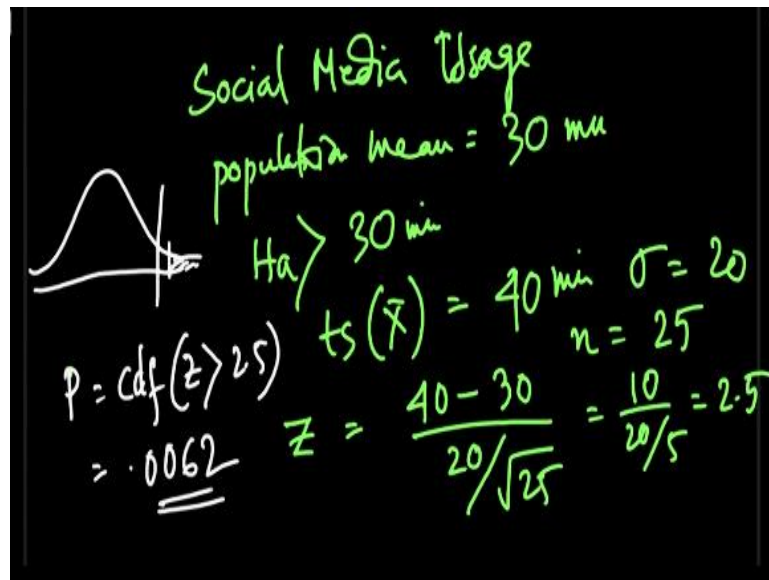
So the depending on the position of the z statistic, we are calculating the P value. It is as simple as that. So the whole area under the curve is representing the all possible you know possibilities where the z can lie and P value is actually representing the area under the curve depending on where the z is actually lying, okay. So that is what the P value is.

And it is quite obvious because if the P value is here, so that is basically it means the z statistic is lying pretty extreme from the population parameter. And when it is lying pretty extreme to the population parameter, we want to actually say that the sample is

perhaps not representing the population, right? So we reject the null hypothesis. So that is basically the idea of P value.

All right, so with this, we will actually do one example. Just one example of P value. And let us say this example is about, let us say social media usage, okay?

(Refer Slide Time: 10:47)



Social media usage. Let us say social media usage of for the students in IIT Kharagpur. And let us say we have national average, let us say population mean. So let us say people usually use around 30 minutes per day, you know they use social media, okay. And let us say so my alternative hypothesis is that the social media usage is let us say more than 30 minutes.

So if I want to see if the students in IIT Kharagpur actually they use social media more than 30 minutes a day. And what I observe when I actually collect the data I observe the test statistic of \bar{x} for IIT Kharagpur, I see that it is actually 40 minutes okay. And let us say we have given sigma is 20. And my number of observation is 25.

So I have taken data from 25 individuals and I found that the actually average is 40 minutes. So now I have to sort of decide whether I can reject the null hypothesis that the population mean is higher than the, if the population mean is equal to 30 minute sorry, the mean usage of social, mean social media usage is 30 minute okay. So to do that, I will what I need to do first I need to have my z statistic.

My z statistic is going to be $\bar{x} - \mu$ by σ , which is $\bar{x} - \mu$ and by σ is $\frac{\sigma}{\sqrt{n}}$, that is 25. So that means $\frac{10}{20 \times 5}$. So that is 2.5 okay. Now for a z of 2.5 for a one-tailed test, we have the value. So essentially I have to, so I have to get the P value and P value is nothing but I will use a different color. So the P value is going to be the CDF, CDF that z is more than 2.5.

So if z is more than 2.5, so you remember the distribution, if z is more than 2.5 here, so I am going to get this area. So I have to see if this area is falling, you know right or left to the level of confidence, alright? So if I actually get this number from z table standard, z table, we can actually see that. So this is what we will get, 0062. Now 0062 is much less than a 5% significance level or 1% significance level.

So essentially, we reject the null hypothesis. So essentially, we mean the students what is can conclude the students in IIT Kharagpur are actually using spending more time on social media than the population, you know than the population. So that is basically the you know concept of P value and that is basically one application of P value.

But going forward, we are going to see many usage of P value and it will always important to sort of remember where from the P value has come? How you know the concept behind deriving P value instead of just mechanically you know kind of saying that P value is less than so and so. So I reject the null. It is we can actually do a better job with P value. So with this, we end this lecture here.

And going forward we are going to talk some other concepts about hypothesis testing.