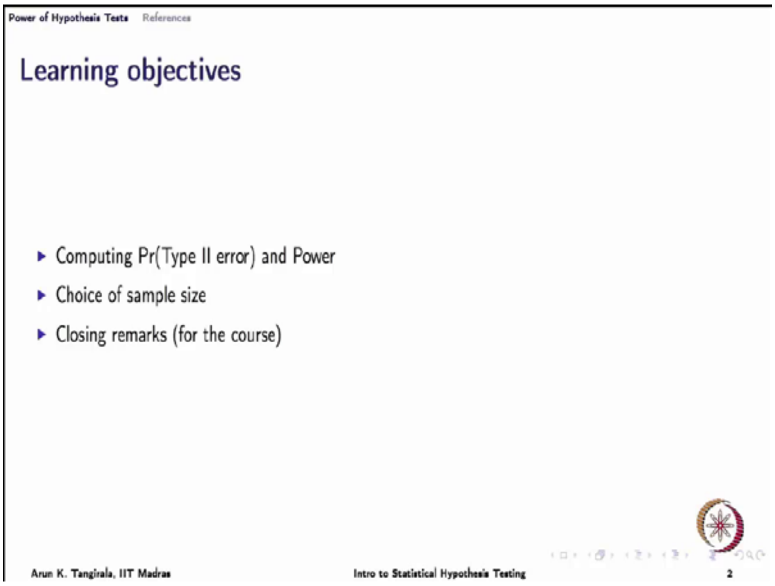
Introduction to Statistical Hypothesis Testing
Prof. Arun K. Tangirala
Department of Chemical Engineering
Indian Institute of Technology, Madras

Lecture - 18
Power of Hypothesis Tests

Hello friends, welcome to the closing lecture on the course on Introduction to Statistical Hypothesis Testing. We have, in this course, gone quite a way in learning what is hypothesis testing and how does one conduct setup and conduct hypothesis test and the concepts associated with it. Of course, the course has been fairly introductory, but hopefully has covered at least the salient concepts that you are looking for. In this lecture, we are going to talk about what is known as the power of hypothesis test.

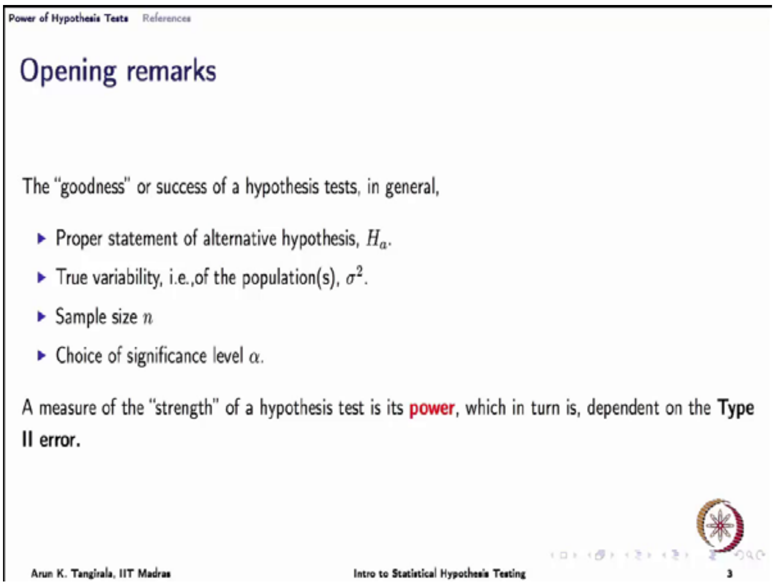
(Refer Slide Time: 00:52)



We have spoken about this before when we were introducing certain terminology in hypothesis testing, where we talked about type I error, type II error, and so on. And in that cont, we introduce this term, power. So, we will look at this concept more in detail. Of course, for this specific case of one sample test for mean, the entire lecture is restricted to that, and extensions to other hypothesis tests are pretty straight forward, in the sense of concepts the expressions may be a bit more involved for the other cases, but if you have understood how to compute the power, and of course, the type II error or the probability of type II error for the one sample test for mean, you are pretty much in a good position to do the same for other kinds of hypothesis tests. And typically, one asks in any experiment for involving statistical analysis and design as to, what is a choice? What should be a good choice of sample size? In order to draw meaningful and reliable inferences, and also in the context of hypothesis testing to sensitize it to the smallest deviation, for example, of the truth from the postulated value.

Remember, in a hypothesis tests, I postulate a certain value for the parameter and the truth could be something else. I would like to collect enough data, for example, that is even able to detect the smallest shifts or the smallest deviation of truth from the postulated value. We will look at that question and we will also look at the related question as to for a given sample size and significance level and power, what is the smallest level of deviation that one can detect. So, there are both these question that we look at, again restricted to the one sample test for mean case. And then, of course, make a few closing remarks not only for this lecture, but also for this course. So, let us get going.

(Refer Slide Time: 03:11)



To begin with remember and recall that the goodness or the success of a hypothesis tests actually depends on several factors starting from the statement of alternate hypothesis to the choice of significance level. So, we have seen that it is important to choose the alternate hypothesis carefully depending on the problem statement, and of course, it is more or less fixed by what you are searching for. So, there is not much freedom there as far as the user or the experimental analyst is concerned. And the next factor that affects the outcome or the decision of a hypothesis test is the population variability. Of course, variability may not always be the factor, but even if you are looking at two sample cases whether you are comparing means, proportions and so on. Variability plays a significant role either directly or indirectly.

Again, from an experimentalist view point, from an analyst view point, there is hardly any control on the true variability. May be one could actually choose, for example, a nice sensor with high precision if the source of variability is due to the sensor, but by and large, the variability that we are talking about is not of the sensor but of the process itself, which we do not have much control over beyond a certain point

And then comes the sample size, which is perhaps the factor that is largely in the hands of the experimentalist, but sometimes may not be because it depends on how long or how much effort one has to put in, in collecting a single observation and so on. But let us not worry about those cases. By and large, one has pretty good control on the sample size and what we would like to know is the relation between sample size and the success of a hypothesis test. In general, if you look at estimation theory, one looks at the relationship between the sample size and the error in parameter estimates. Here, we are looking at sample size, the relationship between sample size and the power of hypothesis test; they are related but the expressions are quite different. Of course, we are not going to talk about the problem of parameter estimation; we are talking of the problem of hypothesis test. So, the expressions that we look at will allow us to determine an appropriate sample size for a given significance level, power and also the deviation of the truth from the postulated value, if any.

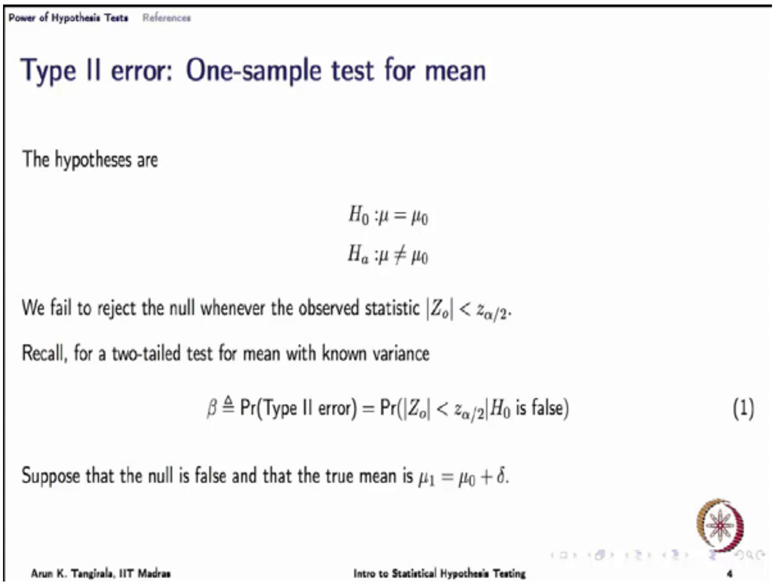
Finally, we also have the choice of significance level. If you recall, one of the things that we said is that this alpha is nothing but your type I error, and as it stands, there is quite a bit of freedom on the user's part; but then, remember that there are 2 kinds of risks involved in hypothesis test. The alpha corresponds to one kind of risks, which is a risk of rejecting H naught when it is true and that we call as a type I error or the probability of the type I error, strictly speaking. Then, there is this other kind of risk, which is the probability of type II error when H naught that is the null hypothesis is false and we fail to reject the null hypothesis; that is, the person who is standing in the court, who has been accused, he is indeed guilty but the person is acquitted. So, that is the other error or the other type II error and the probability of that is what is called beta.

So, if as much as it appears that I have freedom in choosing the significance level and I would ideally like to minimize the type I error, but if you recall, we made a statement that life is not that easy, while we meddle with or interfere with the significance level here. We are also in parallel altering the probability of type II error, and unfortunately, in an inverse way; in the sense that if I actually; it is not inverse mathematically but qualitatively; if I decrease the significance level, that is if I try to decrease the type I error, I am devising a test. I want to make sure that the probability of type I error is low, then somewhere I am increasing the type II error, probability of type II error. So, we are going to see that today quantitatively. Earlier, I had promised to you that we will show you as to how the type I and type II errors are related and this is a lecture in which we will do that.

The focus of today’s lecture, as I mentioned early on, is the power of hypothesis tests which determines how strong the test is in terms of rejecting H naught when it is false, because that a really is a good way of what do you say, characterizing a hypothesis test, because the default situation is H naught. Now, the data is the evidence for determining whether the H naught should be rejected or not, but apart from the data there is a test statistic, which is like, as I been saying always the magnifying lens for the detective to search for the evidence. And this test statistic actually takes birth from the estimator itself. So, in some sense, we want to have a test statistic that is quite sensitive to the smallest deviation of the truth from the postulated value; that is even when H naught is false for the smallest deviation my hypothesis test should be able to detect that. And that is why we base the definition of the power of the hypothesis test on the probability of type II error which we denote as beta.

As I mentioned early on, we restrict ourselves to the one sample test for mean and entire discussion is built on this, but the concepts are quite universal for other hypothesis tests as well.

(Refer Slide Time: 09:49)

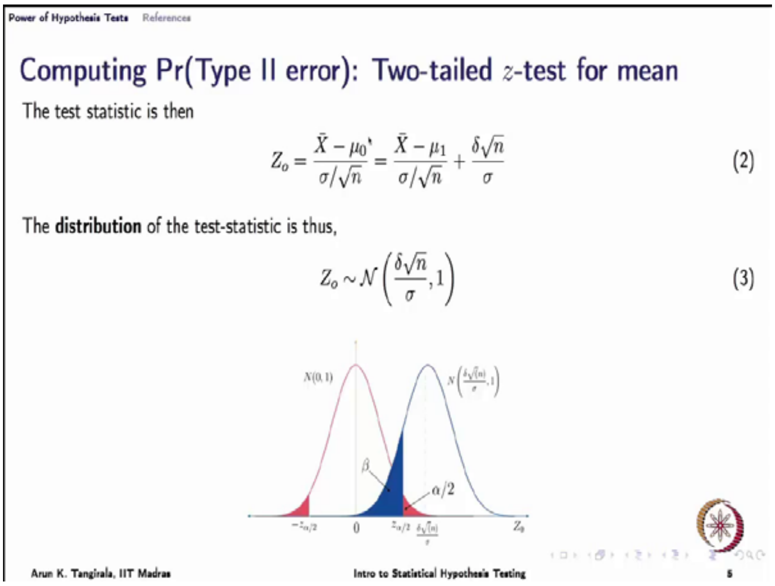


So, let us now look at the one sample test for mean and talk of the probability of type II error. To recall, the hypothesis tests for the one sample test for mean involves null hypothesis mu equals mu naught and the alternative hypothesis being one of the 3 possibilities. Again, we will initially leverage our discussion on the two-sided test, but then, we will also talk about the one-sided tests that is the upper and the lower tailed test. So, suppose I am looking at a two-tailed sample test for mean and where mu naught is the postulated value, then we fail to reject the null hypothesis whenever the observed statistic Z o is less than Z alpha by 2, where alpha is a significance level because Z alpha by 2 is the critical value.

And Z is our usual standard, the random variable that follows a standard Gaussian distribution. Now, if you recall, for a two-tailed test for mean with known variance beta, which is a probability of type II error is the probability of the observed statistic falling within the acceptable region even as H naught is false, because that is what essentially type II error is; we fail to reject H naught even as H naught is false. So, when do we fail to reject the null hypothesis? When, the observed statistic falls in the acceptable region. When does the observed statistic fall in the acceptable region? When, Z o, that is the magnitude of Z o, is less than Z alpha by 2. So, beta is a probability of such an occurrence.

You have drawn the samples from a distribution with mean different from the postulated value. And, let us call that value as mu 1; that is mu 1 is a truth, mu naught is a postulated value. Let the mu 1 be different from mu naught by delta; delta could be less than 0 or greater than 0, that is not an issue here. So, the data comes from mu 1; postulated value is mu naught. And, now we need a way of computing this beta. It is pretty straight forward although a few steps are involved. Most of the steps are just common sense. We just go by this definition and equation one, and derive the expression for computing beta.

(Refer Slide Time: 12:15)



Now, let us get started. First of all, the test statistic is Z o which is X bar minus mu naught by sigma over root n, which we rewrite in terms of mu 1 and delta. So, in place of mu naught, we write mu 1 plus delta and that is very simple algebraic operation there. So, Z observed, actually consists of 2 parts: X bar minus mu 1 by sigma over root n plus delta root n by sigma. Now, when this is a situation; what is a situation? That X naught is false, that is a true mean is mu 1; the distribution of the test statistic actually follows. Still follows the Gaussian distribution, but with a different mean. What do we mean by different mean? If delta was 0, that means, if mu naught was the same, as a truth we very well know that Z observed has a mean 0; it has a standard Gaussian distribution with mean 0.

But now the mean is mu one instead of mu naught; therefore, the Z observed has a mean, and the mean shift is this delta root n by sigma. In other words, Z observed can be thought of as a sum of two variables: the first variable is a random variable; delta root n by sigma is not a random variable; delta root n by sigma is simply a constant; we may, we do not know that because delta is usually unknown, we assume sigma is known here - in this one sample Z test two-tailed. Therefore, the first one is a random variable, the second one is a constant, and the second one essentially causes a mean shift in Z observed.

Remember, since x bar which is computed from the data, where is the data coming from? From a population with mean mu 1, naturally, this first random variable has a mean Zero. That is, if I compute expectation of Z observed, it is the sum of these 2, expectation of the sum of these 2 terms, the expectation of the first term is 0, because x bar is sample mean, it is an unbiased estimator, and x bar is computed from data which comes from the truth; that means, X bar should also have a mean mu 1. Therefore, the first term has a mean 0; the second term is simply a constant; therefore, expectation of Z observed is delta root n by sigma. And because delta root n by sigma is a constant, it does not really contribute in any way to the standard deviation of the Z observed. The only term that contributes to the standard deviation is the first term, and we know, once again, that since x bar is coming from a population with mean mu 1, this standardized random variable, standardized sample mean, has a variance of one or standard deviation of one. In other words, this difference in the truth and the postulated value has only caused a mean shift in the observed statistic, but otherwise a standard deviation remains the same.

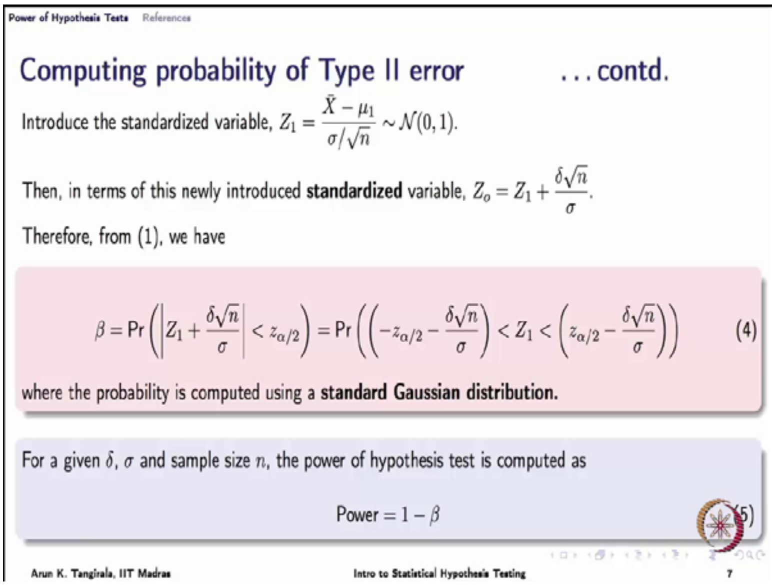
The figure here which has a graphical representation of what we just discussed, tells us the same thing. Essentially it says that both, while the red line here corresponds to what we assume based on the postulated value; the blue line here corresponds to the reality that is in reality Z observed is actually following the distribution shown in blue color.

But when we conduct the hypothesis test, we assume that the Z observed follows the distribution shown in red. Why do we assume that? Because we hold the null hypothesis to be true; whereas, the reality is it is actually following the distribution in blue, but if you look at both these graphs they are the same in terms of spread and distribution, except that they have different means, they are centered; one is centered around 0 - the red one is centered around 0, and the blue one is centered around delta root n by sigma.

Now, beta, which is a probability of type II error, is this blue shaded region; that is the probability of Z observed falling between Z alpha by 2, minus Z alpha by 2 and Z alpha by 2. Of course, this graph is only been shown here for delta greater than 0.

But you could have this graph here; there is a blue one the truth being to the left of the red line that is delta could have been negative as well. We are just showing qualitatively how the truth would look like when delta is greater than 0. So, hopefully now you understand what is the beta exactly. Of course, now what remains to be understood is how to compute beta.

(Refer Slide Time: 17:23)



So, in order to compute the type II error we are just one step away, in fact. For this purpose, let us introduce the standardized variable Z 1. So, earlier, we had Z observed as a standardized variable which used mu naught that is the postulated value as a reference. Now, Z 1 uses mu 1, the truth, as the reference. So, Z 1 is x bar minus mu 1 or sigma or root n. And, since x bar is computed from data which is coming from the truth that has mean mu 1, naturally from the property of sample mean, we know that Z 1 follows a standard Gaussian distribution.

So, all that is required is now to rewrite Z observed in terms of Z 1. And we know, from the definition of Z observed and that of Z 1, they both are related as Z observed equals Z 1 plus delta root n by sigma. So, we just take this expression for Z observed and plug into our expression here for beta and here in place of Z 1. Therefore, we have Z 1 sorry;

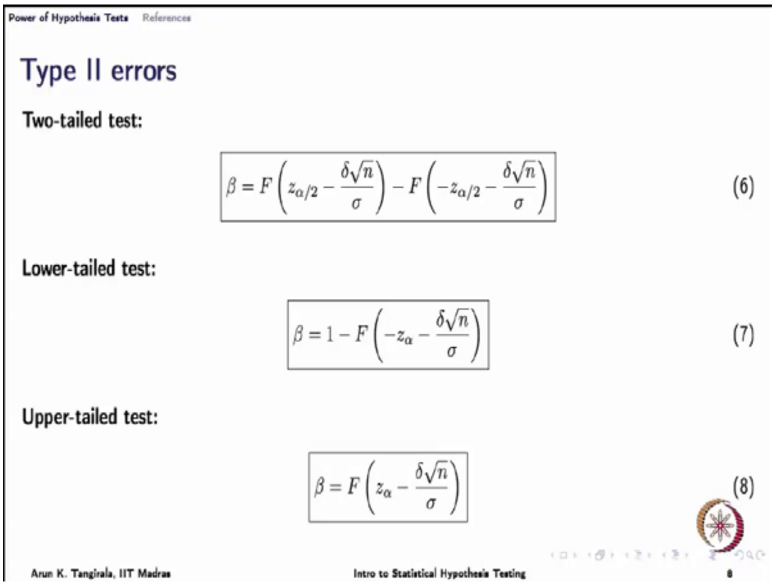
Z observed we have Z 1 plus delta root n by sigma. So, beta is a probability that the magnitude of Z 1 plus delta root n by sigma is less than Z alpha by 2. And which now, upon further expanding and rewriting in terms of Z 1, we have beta as a probability that Z 1 lies in the interval with the left bound as minus Z alpha by 2 minus delta root n by sigma and the right bound of the interval is Z alpha by 2 minus delta root n by sigma; that is it.

So, do we have everything to compute beta, let's take a quick check. Here, we know we need to specify alpha, once alpha is specified Z alpha by 2 is known. We need to specify sigma that is known and n is the sample size, we know it. Delta is the difference between the postulated and the truth. So, that is also known. We have everything to compute the left and right bounds, all one does need to know is go to a standard Gaussian distribution table, if you do not have access to a computer use a standard Gaussian distribution chart, and look up the probability of a random variable following a standard Gaussian distribution lying in this interval; that is all. Of course, if you have an access to a computer, like we do in the courses, in the assignments and so on, you can use the p norm to compute the probability; that is all.

Now, as you can see the procedure to compute beta is pretty straight forward. We have demonstrated this for the two-tailed test. You can also repeat this exercise. I suggest that you do it for the upper and the lower tailed test. Everything remains the same; Z observed remains the same, Z 1 remains the same, except that instead of this beta being the probability of Z observed being in interval minus Z alpha by 2 and Z alpha by 2, for the upper tailed test, for example, you would have beta as the probability of Z observed less than Z alpha. Of course, given that H naught is false and likewise, for a lower tailed test probability that Z observed is greater than minus Z alpha, that is it. Otherwise, everything else remains the same.

With that in mind we will move on to now the definition of power. Power is nothing but 1 minus beta. For your convenience, what we have done here is we have summarized the expressions for the power of the hypothesis tests for different scenarios.

(Refer Slide Time: 21:14)



That is for the two-tailed test, the lower-tailed test and the upper-tailed test. The expression for the two-tailed test simply follows from the expression that we had in equation 4. While for the expressions in 7 and 8, you can easily derive by going back to this derivation for the two-tailed test. And, all you have to do is, replace here in expression 4. The probability of the mod Z observed less than Z alpha by 2 with probability of Z observed being greater than minus Z alpha for the lower-tailed test or Z observed being less than Z alpha for the upper tailed test. Once you do that, and remember that we are working with the cumulative distribution functions, which give the probability of the random variable taking on any values from minus infinity to a specified value. You get for the lower-tailed test this kind of an expression, 1 minus f of minus Z alpha minus delta root n by sigma; whereas, for the upper-tailed test you would get f of Z alpha minus delta root n by alpha.