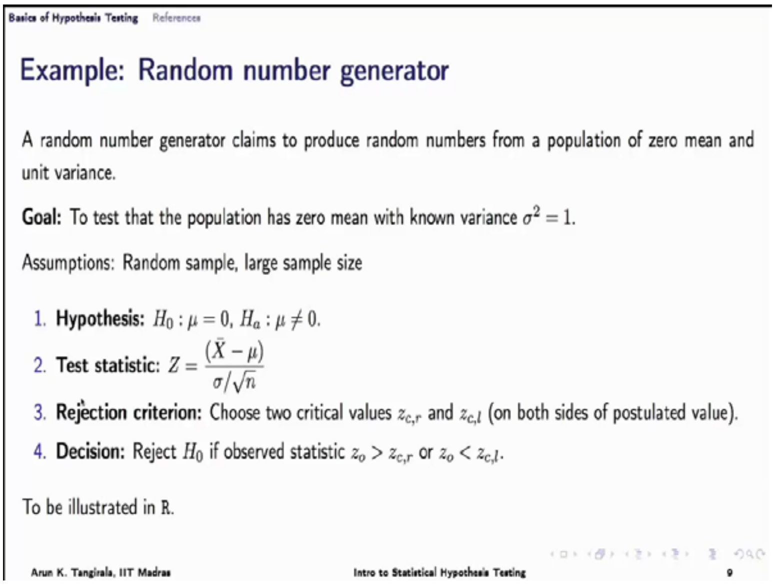
Introduction to Statistical Hypothesis Testing
Prof. Arun K. Tangirala
Department of Chemical Engineering
Indian Institute of Technology, Madras

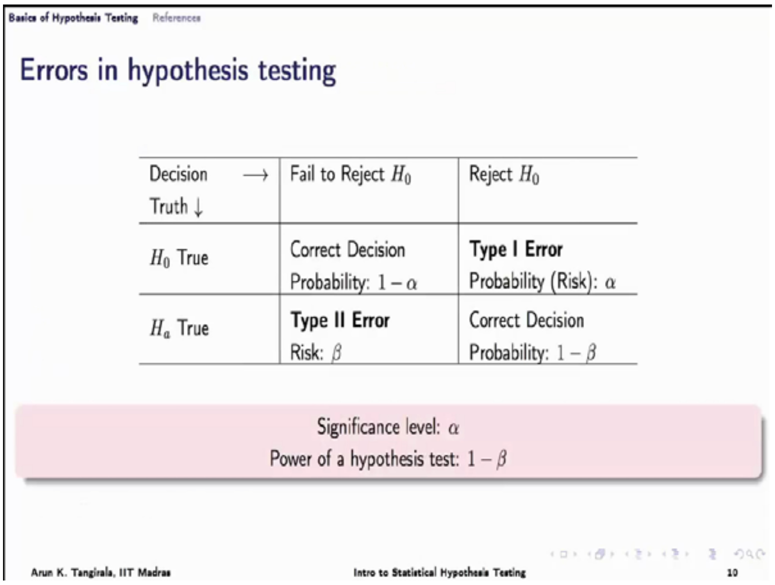
Lecture – 10
p-value

(Refer Slide Time: 00:01)



Alright, so now, with this illustration we can discuss certain theoretical aspects of these errors in hypothesis testing.

(Refer Slide Time: 00:02)



So, as I said earlier there are 4 possibilities as listed in this table, the columns contain the decisions and the rows contain the truths. Suppose, I look at a first row were H naught is true, there are 2 possibilities. I will fail to reject the null hypothesis or I will reject the null hypothesis. If H naught true and I fail to reject the null hypothesis, excellent that is a correct decision no worries, but as we have seen in the example, it is not necessarily that I will make this correct decision all the times, which means the probability of making a correct decision when H naught is true and failing to reject H naught is not 1 it is less. Then 1 and in fact, it is 1 minus alpha because the probability of rejecting H naught when, H naught is true is alpha there only 2 possibilities. Essentially; what you are doing is you are partitioning the space of possible values for the statistic into 2 parts acceptable region and critical region that is why hypothesis testing is also like the classification problem.

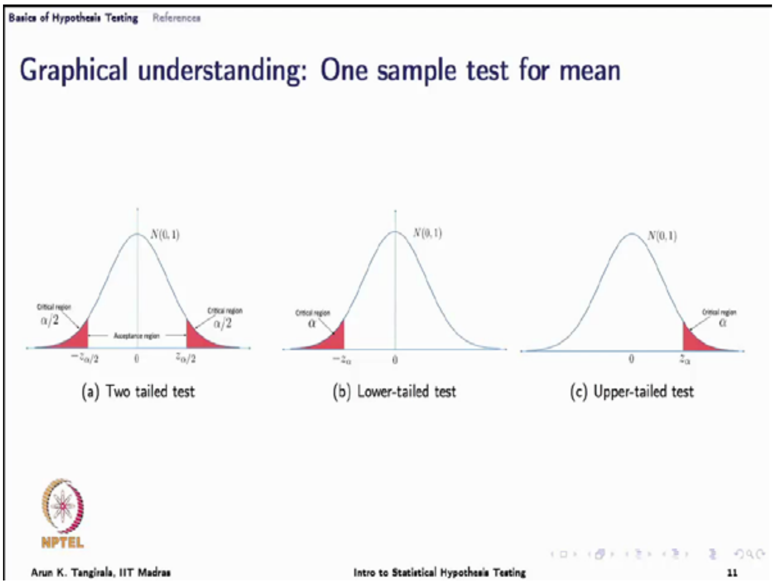
You can think of it that way, this probability of rejecting H naught when H naught is true is also a risk. Why is it called a risk? Well, let us go to the example of this jury trial which we talked about in 1 of the previous lectures. In a court of law the null hypothesis is always that the accused his innocent. That is the null hypothesis that the judge keeps in mind and the lawyer also keeps in mind and it is a prosecutors burden to provide sufficient evidence to show that to prove that the null hypothesis has to be rejected. Now, suppose this person is actually innocent and the judge proclaims the person to be guilty, there is a risk associated with it. In many different ways, it could lead to psychological problems for the person it is a trauma of course is no doubt. It could lead major problems for the society and so on. So, it is a risk and it is a risk for every kind of hypothesis testing problem. So, it is one type of risk, the other type of risk that we encounter is in the second case when the alternate hypothesis is true; that means H naught is false and if we do not reject null hypothesis this is the case when the person is indeed guilty and the judge acquits this person. That is also risk and we know why, because this person can actually go and create hawk again or commit crimes again. This risk is called type II error and we denote this as beta and earlier when I said when we try to minimize the type 1 error as much as we want to we do end up unfortunately increasing the type II error.

It is not that alpha plus beta is 1, please do not get into that but there is a relation is kind of a water bed effect if I try to decrease alpha the beta will increase and vice verse. If I want to decrease beta then I have to sacrifice alpha. Of course, we have the 4 possibility that which is a strongest one; that is the person he is guilty and the enough evidence has been provided and judge finds the person guilty. That is alternative hypothesis is true, H naught is false and we have rejected H naught that is a correct decision. So, there are 2 correct decisions and 2 risks, but the strongest one is the fourth one. That is fourth possibility that we discussed that is, correctly rejecting H naught when it is false. Why? Because the default as we have said, is always a null hypothesis and when a hypothesis test has managed to detect the falseness of H naught with the evidence that has to be provided it is a very good hypothesis test and therefore, the power of hypothesis test is characterized in fact, the probability of making this correct decision that we reject H naught when it is false.

So, the by the relation in the table the power of the hypothesis test is 1 minus beta. However, when it comes to testing the hypothesis we ask whether we specify alpha do we specify beta. In the earlier illustration, what we did is we specified the critical values and computed the alpha but remember I said there is a second approach to hypothesis testing. Where we specify the errors and then determine the critical value. Now, the question is should I specify alpha or should I specify beta. If you can quickly understand from the table alpha is associated with rejecting H naught when it is true whereas, beta is associated with failing to reject H naught when it is false. Now; obviously, it is easier to calculate or specify alpha then beta because, to specify beta I need to know the truth because it is the probability of failing to reject H naught when it is false. When H naught false there are several possibilities we do not know what the possibilities is.

All we are saying for example in a two-tailed test, we are saying mu is not equal to mu naught. When mu is not equal to mu naught, there are several possibilities. So, to compute beta I have to again postulate some truth that is different from what is postulated in x naught therefore, it becomes difficult and we will see later on in a lecture in the one of the final lectures. What is the relation between beta, the sample size and the actual truth; there is expression that allows us to calculated it but it is a bit more involved as I said just now, by specifying beta we are actually implicitly saying the truth is something different mu naught instead of mu naught it may be mu 1 but that is not what we get into, what we want get into is in the context of null hypothesis that hankered around the null hypothesis we want to specify the type I error because that does not make any assumption about the truth if H naught is false. In summary specifying alpha is much easier then specifying beta and that is why in practice you see this, significance level being mentioned in every hypothesis testing problem, right. That is the story of the significance level for you.

(Refer Slide Time: 06:58)



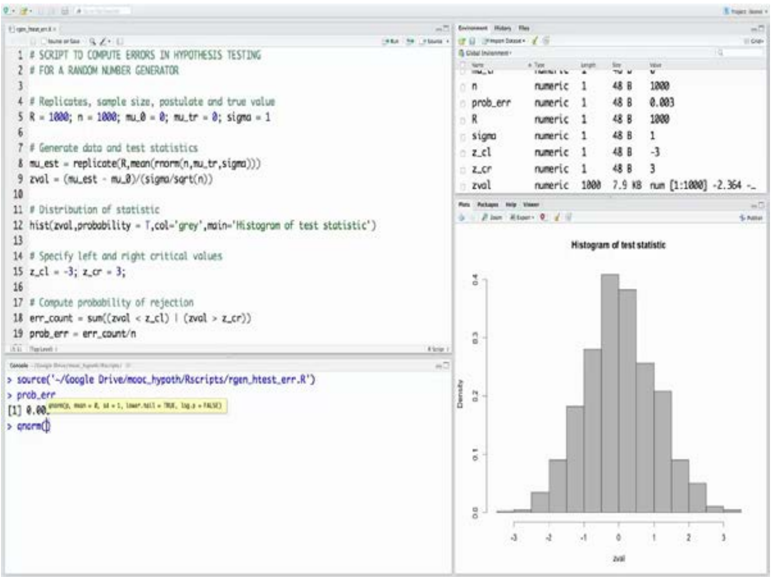
And to further reinforce what we are just discussed we go through a graphical understanding again a caution, the curves that you are seeing are for the sampling distribution of a test statistic that follows a standard Gaussian; which could be the case of sample mean or may be you know sample proportion and so on. So, the attention should not be paid so much to the shape of the curve as much as it should be paid to the critical region. What we are saying is in a two-tailed test you will have critical regions to the left and right of the postulated value because of the nature of the alternate hypothesis. Where as in a lower or a left sided lower-tailed, or left sided test the critical region is to the left of the postulated value and in the upper-tailed test naturally the critical region is to the right of the postulated value.

Now, one thing that should be remembered is this area. In fact, there are 2 points; the area corresponding to the red shaded region is what an alpha. So, let us look at the two-tailed testing, we have two critical regions shaded in red and we have indicated the areas to be alpha by 2. Is this always the case well no it is the case only when this sampling distribution of the test statistic is symmetric. That is for the case of sample mean because the sample mean on the other hand, if we have the case of the testing of variability where the distribution is chi square the density curve we know is asymmetric in which case the critical regions are not going to be symmetric that is they will not have the same area the probabilities are not going to be the same to the left and right of the postulated value. How does it matter? Well, it does matter because the procedure that we shall follow in hypothesis testing as we are outlined one approach is to specify the critical value by the user but then you will have to keep computing the type I error each time.

And, the other procedure that is followed that is practiced in a standard way is to specify the type one error. So, when I specify the type I error which is alpha, I then fall a compute the critical value remember this, we have here z alpha by 2 and minus z alpha by 2. What are these critical values? How do you compute this? Well, we compute this with the help of the tables that are probability distribution tables, that are given in a book or more modern way is use software such as r to determine this so-called quantile corresponding to the probability.

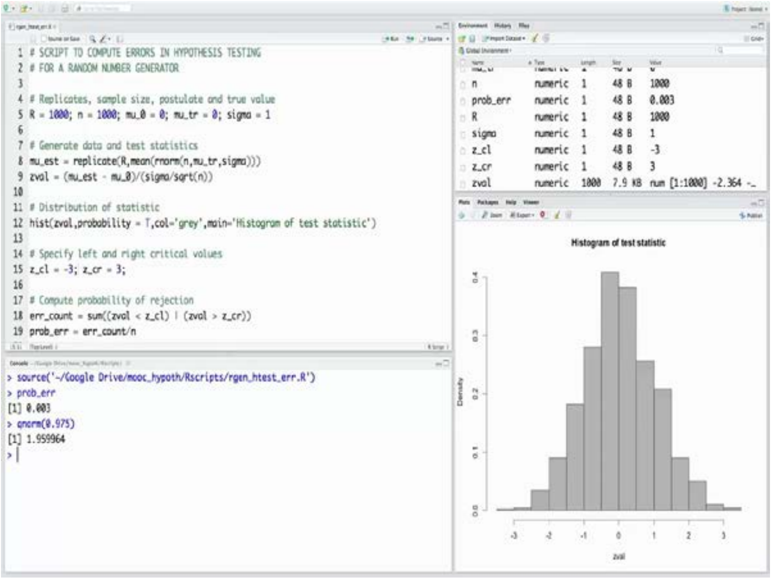
As an example, suppose alpha is 0.05 and the hypothesis test is a two-tailed test and we assume that the sampling distribution of the test statistic is standard Gaussian as shown in the figure. And, I want to figure out what this z alpha by 2 is what we do is we go to r and find this quantile, what quantile do I want to find what is the quantile give me in r? It tells me what is this value for given probability remember, if I specify this value on the x axis I can calculate the probability that is what p-norm for example would do. Q-norm will tell me the reverse that is I specify the probability and it tells me the value on the x axis, which probability should I specify; the probability that the z statistic will take on any value from the left extreme to this z alpha by 2. That is how q-norm is written, for in any other software package it may be different. So, you have to watch out. As far as r is concerned q norm will give me the value on the x axis for the probability that this random variable takes on any value from the left extreme to up to that value. When alpha is 0.05 alpha by 2 is 0.025; that means these 2 shaded regions here have an area of 0.025. What I am going to do is to determine z alpha by 2 I got to q-norm.

(Refer Slide Time: 11:42)



And, ask for the complimentary probability because I know 0.025 is the area of the red region and I just know said q-norm looks at the probability to the left from left extreme to z alpha by 2 and that obviously is 0.975.

(Refer Slide Time: 12:05)

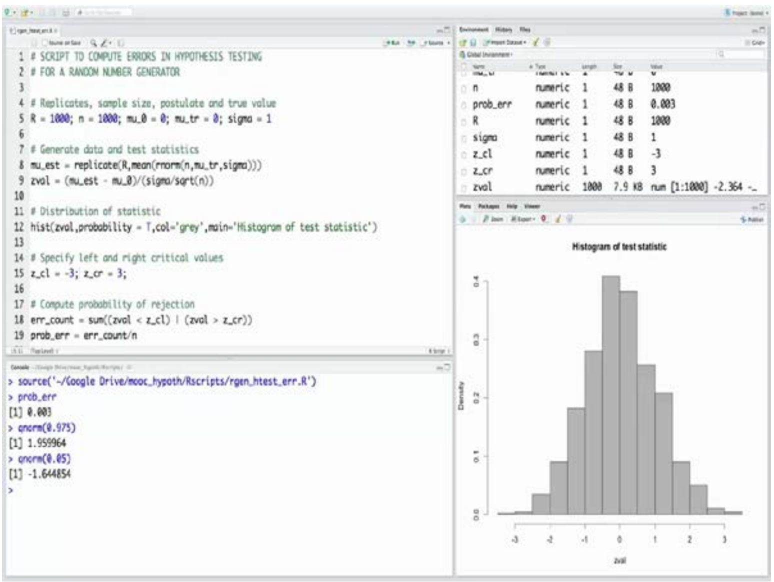


So, we asked what is the quantile corresponding to this and there you go I get 1.96 roughly which we know 95 percent probability is what essentially is corresponding to the acceptance region since it is a Gaussian distribution minus z alpha by 2 is minus 1.96. If it was a chi square distribution I have to calculate the right and left critical values separately they are not going to be simply the left critical value is not simply, the right critical value with the sign reversed. We will have to calculate them separately. We will discuss that in one of the next lectures were will talk of hypothesis test on variability. Let us get back to the discussion now.

So, what we have learnt is for a given alpha or a specified alpha, I can compute the critical values and the advantage is now, I know the type I error that is probability of rejecting H naught. When it is true of course, the question that comes to mind is can I calculate beta given alpha. Yes, I can but, the calculation is a bit more involved we will not get into that this movement will only qualitatively discuss what is beta shortly in the next slide. Let us complete this discussion on determining critical values by specifying the type I error. So, it is a two-tailed test the critical value is the z alpha by 2 alpha the alpha by 2 corresponds to the area of the shaded region on the right in a lower-tailed test as you have seen the critical value is to the left and the subscript there is alpha on z, we have minus z alpha clearly telling you that the critical values going to be of negative sign.

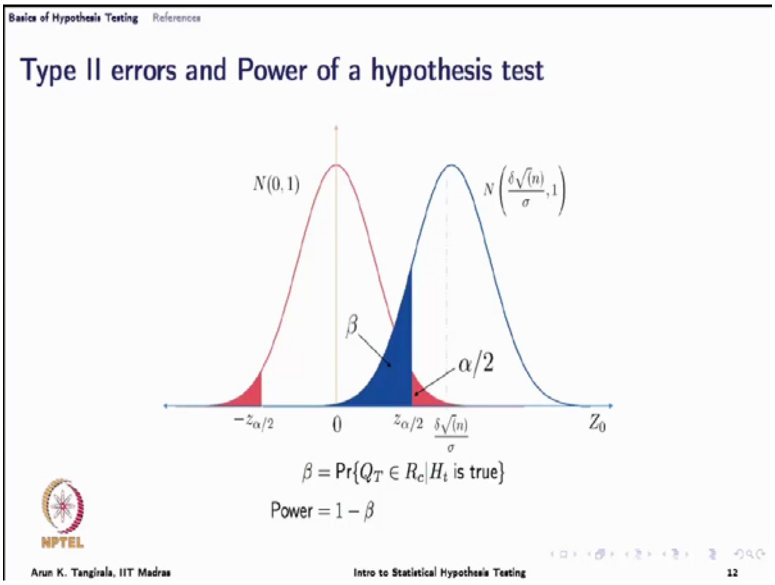
In this case, the critical region is completely to the left because we are saying, when does this arise when the alternative hypothesis is that of the type theta is less then theta naught or mu is less then mu naught; that means, when the test statistic false to the left of the extreme value. That I am willing to tolerate which is the critical value then, I reject the null hypothesis saying that, No, if the null hypothesis is true it is very unlikely that I can get an extreme value such as this. Now once again, here I can calculate the critical value but the only difference between the two-tailed and the lower-tailed test is that, the determination of the critical value in this case the critical value can be computed again by going back to q-norm, but now the probability that I am going to specify is, can you guess? Well, that is 0.05 because that is alpha is what is the error specification.

(Refer Slide Time: 15:05)



Let us ask, what is that? Of course, most of us know the answer. Now, here I do not subtract because subtract it from one. Like I did earlier because, we are looking at a lower tailed test that something you should remember and there you go you get minus 1.65 roughly, right and you should guess. Therefore, that in an upper-tailed test the critical value which is going to lie to the right of the postulated value is going to be plus 165 the rest of the story is the same. Now, we know for a given alpha how actually we can determine a critical value.

(Refer Slide Time: 15:43)

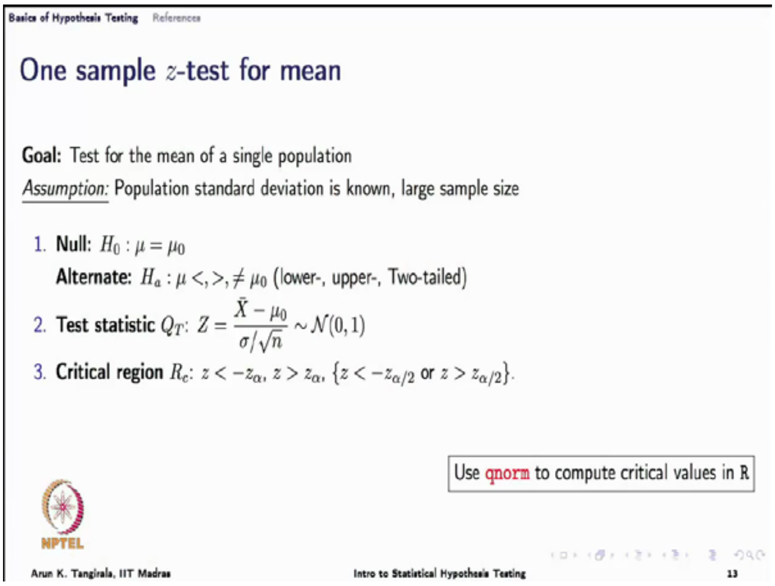


Earlier, we raise this question can I determine beta that is type II error and subsequently, calculate the power of a hypothesis test. Yes, we can. I am not going to show you how to calculate that now. However, I am just illustrating what beta is graphically, we have seen what alpha is graphically and I am illustrating this for a two-tailed test, remember that. So, what is this beta if you recall we are saying that the null hypothesis is not correct is not true the alternate hypothesis is the one that is true. Now, the curve that you see in the red is the PDF of the test statistic assuming null hypothesis to be true and the curve that is colored in blue is the null, is the sampling distribution of the test statistic. When the truth is something else that is, if I had specified the right null hypothesis then the sampling distribution would have been or the PDF would have been what you see in blue. So, beta by definition is the probability of failing to reject H naught. When H naught is not true which means the probability of finding the test statistic in the acceptable region when the fact is that the null hypothesis is not true.

So, I indicated now the Z alpha by 2 we know that acceptable cases are whenever, Z falls to the left of Z alpha by 2 and therefore, the beta is simply this shaded region in blue because, in all these possible values what is happening the null hypothesis, is not going to be rejected because Z is going to be less than Z alpha by 2, but we have to calculate that probability when H t is true; what is H t? The true hypothesis which we do we have not specified. We have specified a different hypothesis in the form of null. So, it is a conditional probability that we have calculating symbolically, we say beta is the probability of Q t naught in fact, not belonging to the rejection region belonging to the actually acceptable region when H t is true. So, there is slight mistake will correct that in the slide for you.

So, probability of Q t belonging to the acceptable region, when H t is true and the power is 1 minus beta what happens the Z alpha by 2 suppose I decrease alpha what happens to Z alpha by 2 it moves to the right as a result of which, beta also increases. If I move increase alpha then, Z alpha by 2 would move to the left decrease in beta. So, that is kind of water by defect but you should not be let to be live that alpha plus beta is one obviously, it is end but what you can see is that alpha and beta tie together in and this illustration nicely brings that relationship therefore, the moral of the story is as you try to decrease type I errors you will invariably increase type II errors that is see fact of life in hypothesis testing, alright.

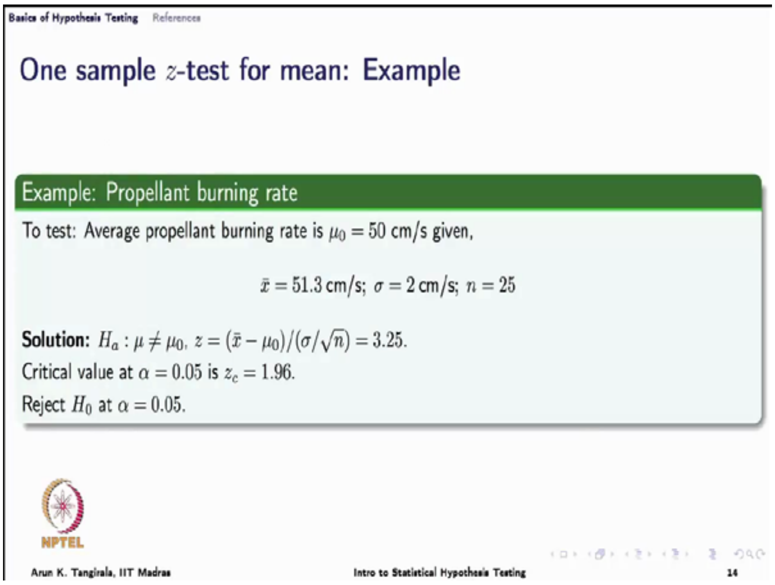
(Refer Slide Time: 19:14)



Now, what we are going to do is we are going to just simply illustrate these ideas and concepts on one example, which is a corresponding to the one population. When we say one sample, it is actually saying one population test for mean. I say is z-test because conventionally people talk of z-test t test and. So, on which I am not so much in favor of the correct way of specifying is what parameter you are testing how many populations and then you can determine whether, it is a z-test or t test depending on the assumptions that you are making but some schools of thought believe that the movement. We say z-test, the assumption such as population standard deviation is known, large sample size automatically quick in will kind of stick to convention in a half-hearted way. So, the goal is to test from the mean of a single population under the assumption that we have just stated and the null is that mu is equal to mu naught.

Alternate could be anyone of this triplet and the test statistic, if we use sample mean is your standardized sample mean which we know under and of course, sigma is known which follows a standard Gaussian distribution and then the critical region depends on the nature of the alternate hypothesis and the specification of alpha and as I have shown you can use q-norm to compute critical values in r.

(Refer Slide Time: 20:44)

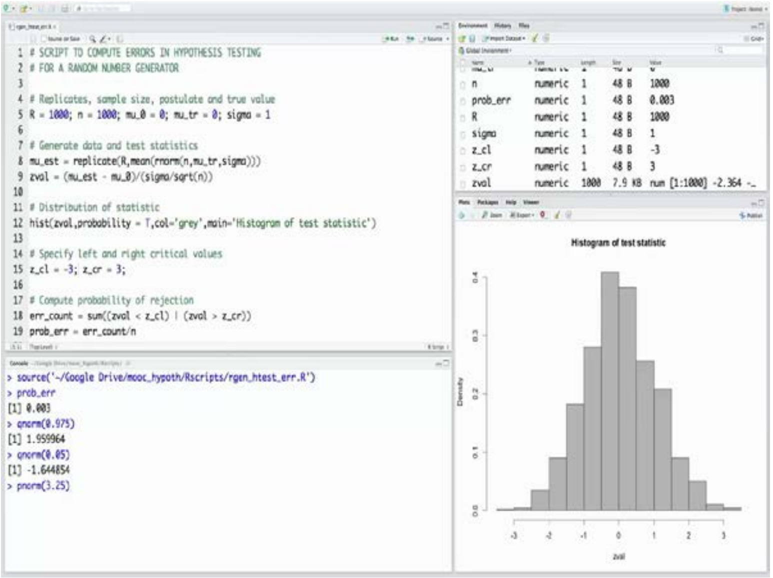


So, let us take this example of the propellant burning rate this is kind of a burning example for us in this course. So, goal is to test whether the average propellant burning rate, you can say is different or not depends on how you put it. Mu naught is 50 centimeters per second and a null hypothesis is that you does not and I am not showing you the data here, I have drawn this example from the book by Montgomery. The sample mean works out to be for a sample size of 25 a random sample of size 25 drawn from the population works out to be 51.3 centimeters per second and assume that the way standard deviation is given as 2 centimeters per second. Which means a variance is for now to work out the hypothesis test we go through the procedure alternate hypothesis mu is not equal to mu naught and we have decided will use the standardized sample mean as a test statistic which works out to be 3.25.

For this example, since it is a two-sided test and we normally specify alpha 0.05, but you can also specify 0.01 that is also a common specification we have just now seen through our previous illustration that the critical values 1.96 and therefore, reject H naught why because the observed test statistic is more extreme then what I can tolerate right it is 3.25 versus what I am willing tolerate is 1.96. Therefore, I reject H naught what this means is I believe that most likely the null hypothesis is not true it is no probabilities that I am calculating here. Probability of H naught not being true and so on, we are not making any such calculations here all we are saying is given the data for this sample size for this specified sigma and specified alpha I find that the observed statistic is more extreme then what I am willing to tolerate. Obviously, if I change alpha to let say 0.101 decrease alpha even to 0.005 or may be 0.001 then the critical value keep shifting to the right as a result of which at some point, I will end up not reject the null hypothesis here.

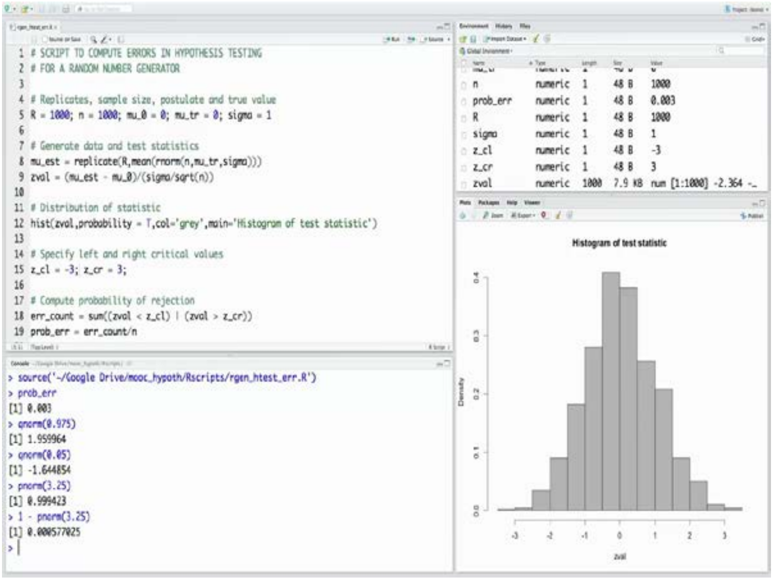
We do not know what the truth is, this is a practical example. So, one question that naturally comes to our minds is, what value of alpha will help me or will end up in the decision say of not failing to reject the sorry to H naught failing to reject H naught that is not rejecting H naught; in other words, how low should the alpha be. So, that the null hypothesis is not rejected. Now, do we calculate this I am sure you must have developed an idea by now think about it is very easy all we have to do is find out what is the probability of the area to the right of 3.25, because think that I had found an alpha such that the null hypothesis is not rejected or you can think of the smallest value of alpha that will reject end up in rejection of null hypothesis. What would have happened? The shaded red region the critical region would have been that alpha and the quantile should have been 3.25. So, in other words what I said earlier is actually should be alpha by 2. In fact, I calculated the area of that is the distribution sorry, density in other words probability to the right of 3.25 and that would be alpha by 2 twice of that would give me alpha. So, we can calculate that in r, right. So, let us do that.

(Refer Slide Time: 24:56)



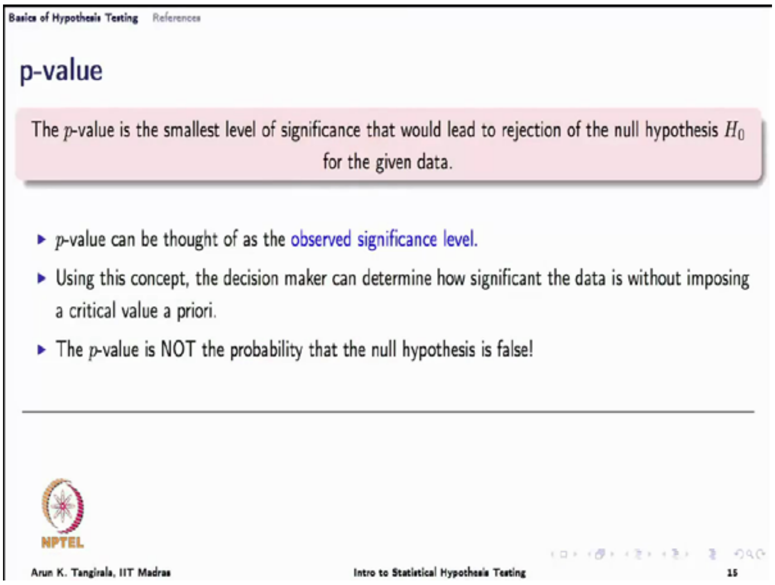
How do we calculate instead of q-norm? Now, we use p-norm and now we asked what is the area to the right of 3.25; because if that I had chosen as alpha by 2 then I would have ended up rejecting the null hypothesis.

(Refer Slide Time: 25:15)



And the p-norm unfortunately now gives me the left probability the complimentary probability. So, what I should be calculating is 1 minus p-norm and that gives me roughly 0.0006, roughly this is alpha by 2. Therefore, alpha is 0.0012, right. So, that is the smallest value of alpha for which the null hypothesis would have rejected any value of alpha less then 0.0012 would have ended up in failing to reject the null hypothesis and that value this the value that we just calculated is called the p-value there are several interpretations to this p-value.

(Refer Slide Time: 26:04)



But, let us look at the definition p-value is a smallest significance level that would lead to rejection of H naught and in for the example, a propellant burning rate example it is 0.0012. You can think of this as an observed significance level. Because we are determining this after the data has been acquired and after the test statistic has been computed. So, it is something observations rather than a pre-specification in that sense the other method where, we pre-specify alpha regardless of the data is called fixed significance level here.

We have an observed or an empirical significance level, there is another prospective to the p-value which is, that is the probability of finding or obtaining a test statistic more extreme then what I have obtained. So, I have obtained 3.25, is there a probability that I can obtain more extreme than this? Yes, and that probability is 0.0012 what is the use of that well that tells me, how extreme the value that I obtained is if look at this example that we just had we said the value is 0.0012; that means, there is a 1.2, 0.12 percent probability of obtaining a value more extremely in this. What is this? Tell me about observed value it is very less likely value when the null hypothesis is true. So, p-value is calculated assuming the null hypothesis is to be true remember that what is that p-value was very high was quite high; that means, that the probability of obtaining more higher values or more extreme values and what I have observed is large when would that happen when the test statistic has fallen in to the acceptable region. In fact, we can use this p-value which is what is done widely is practiced widely I can use this p-value to determine if the null hypothesis should be rejected or not. How I compare the p-value with the fixed significance level that is let say 0.05 and if the p-value is greater than alpha then obviously, it means that the test statistic is an acceptable region and if the p-value is less then alpha then clearly the test statistic falls in the critical or the rejection region and I can therefore, say reject H naught.

So, there now you can actually use p-norm or p t or p chi-square and so on. To compute your p-values we will sight p-values. Whenever, we go through hypothesis test. That is the story of basics of hypothesis testing for you.

(Refer Slide Time: 28:52)

We have in this lecture learnt, what are type I and type II errors; that they are closely related if you try to reduce one the other shoots up. That is a very important think to remember of course, the definitions of type I and type II error; they have gone through a few illustrations and we have also discussed, what is meant by power of a hypothesis test. More importantly, we have also learnt how to determine the critical values when a significance level is specified a priority and we also learnt, concept of value which is the observed significance level which helps me making certain decisions; that means, it tells me how far I could have gone the process could have gone before the null hypothesis is rejected or what is the minimum level of adjustment that I have to make in order actually bring back the null the test statistic into your acceptable region if the null hypothesis has been rejected. So, it is quite useful in to the manufacturers or process design people and also let me tell you there is a fair share of criticism of the use of p-value, but then this criticism depends on how you use the p-value, if you truly understand the definition of p-value and where it should be used, then you are you are not subject to that criticism researchers duty is to point out what are all the areas in which limitations of a certain major and what pit false are associated with the use of either p-value or alpha and so on.

So, one has to be aware of this and learn essentially the definition of those measures like p-value or type one error and so on and also know when and how to use them and that what is one of the primary purposes of this formal course on hypothesis testing. What will do in the next lecture is we will go through a set of examples and the hypothesis test only associated with mean.

Remember, we had so many different situations when we talked of sampling distribution of sample mean we will essentially cover all those cases and then and that subsequent lecture will look at the test on variability, proportion, ratios of variability, differences in proportion with that we would have discussed whatever we want to talk about the traditional way of hypothesis testing.

The last unit we will tell us how to conduct the hypothesis test using confidence region concept, confidence interval concept that is a very beautiful concept and in one of the lectures in this unit we will also learn again come back to the hypothesis test and ask what is a relation between sample size and the type two errors in the hypothesis test or the power of a hypothesis test where we learn that there is an inverse or there is a direct relation as we increase the sample size the power of a hypothesis test typically increases, which is encouraging, which means it is telling that collect more and more samples so that you make as less type II errors as a possible.

We will learn those relations and also learn in all of these next lectures. We will learn how to do things in r, will discuss the theory and I will show you how to do it in r and I will give you a list of commands as well then. So, hopefully you enjoyed this lecture. Will meet again very soon in the next lecture.

Thank you.