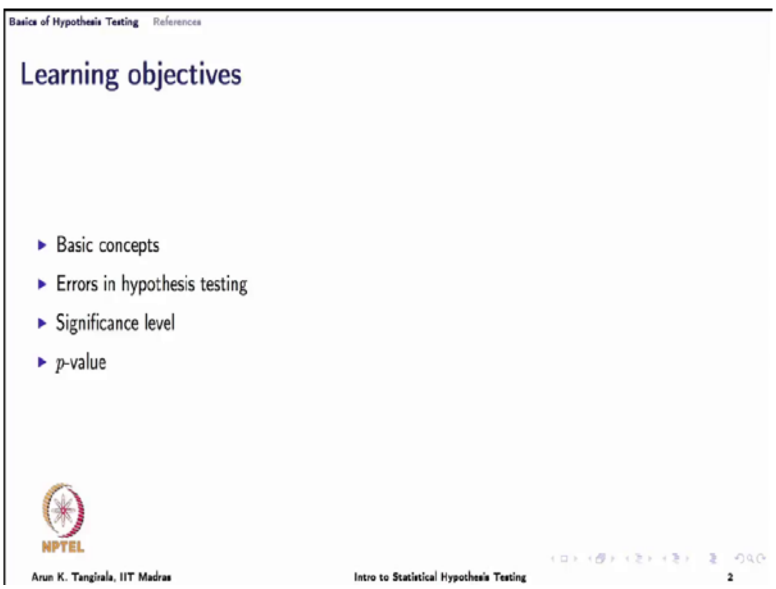
Introduction to Statistical Hypothesis Testing
Prof. Arun K Tangirala
Department of Chemical Engineering
Indian Institute of Technology, Madras

Lecture – 09
Basics of Hypothesis Testing

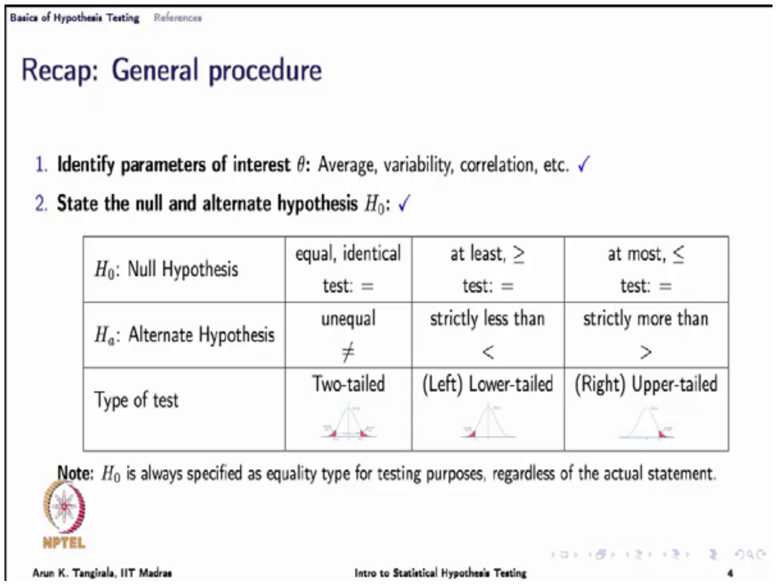
Hello friends, welcome back to the lecturers on Introduction to Statistical Hypothesis Testing. We are into the unit-3. Now, in the first two units, we learned all the theory that is required to take us forward to understand, what is hypothesis testing? Now, we put together all of that and also recall the general procedure to understand the basics of hypothesis testing.

(Refer Slide Time: 00:44)



In this lecture, what will do is we look at certain elementary concepts and learn the definitions of some key terminology in hypothesis testing namely type I errors, type II errors and significance level that one often gets to see in hypothesis testing. It is always a mystery for many as what this significance level is; some people confuse this for confidence intervals and confidence and so on. And then, also one commonly encounters this motion of p-value will also learn that and mostly all of that through an example. And of course, what I will also show you now as the course proceeds will how to do things in hour. For example, we will see the definition of type one error mathematical definition or a statistical definition you can say as well as learn how to compute a type I error in hour. So, it is going to mix of illustrations plus theoretical definitions.

(Refer Slide Time: 01:48)



To recap the general procedure that we learned in the motivation lecture, the first step in any hypothesis-testing problem is to identify the parameters of interest, the parameters that really we want to test. And, in any basic data analysis exercise or even in this course, we are going to restrict ourselves simply to the first moment which is the average, second moment which is variability essentially first and second moments of the PDF or correlation. If it is a bivariate analysis or in a slightly more advanced situation, we are going to look at model parameters such as regression parameters, slope and intercept and that is about it. In a more advanced situation, we actually look at other parameters. These are the parameters of the populations that I have listed here average variability and correlations, so to speak of the truth.

Once we have identify the parameters, obviously, the next step if you recall is to state the null and the alternate hypothesis. Now, this requires some reading of the problem statement, because it varies from problem to problem and it just takes a few practicing on few different problems to learn how to formulate the null and alternate hypothesis. In this table, essentially I have listed all possible scenarios that one could encounter. So, if you take for example, null hypothesis, there are three possibilities one of the equality type like in the solid propellant burning rate example or an at least type where you say for an example in a training method maybe we would say that the training method is effective if the trainees course at least some postulated or pre-specified value. Now, you should also notice that when it comes to testing the hypothesis, the hypothesis of the equality type.

And, if you recall in the previous lecture we went through sample hypothesis testing exercise, where we got a feel hypothesis (Refer Time: 04:06) and the procedure (Refer Time: 04:08) in order to test a hypothesis we assume that the truth is what the null hypothesis states that is, we anchor our results on the null hypothesis. Now, assuming that whatever has been postulated in the null hypothesis correct we compute the probability of a paining this statistic as we have observed or within the vicinity of what we have observed. And, if with that probability, if we find that the statistics falls in a high probability region, then we do not reject the null hypothesis. On the other hand, if the statics falls in a low probability region, we reject the null hypothesis.

Of course, we do make errors in both of this we will talk about that, but as for as a procedure is concerned we anchor on the null hypothesis. And, in order to anchor ourselves and we need an equality type otherwise it becomes very difficult to compute the probabilities, because if I go by the statement at least or at most then there are many possible truths which I have to test simultaneously then it becomes very difficult. So, hypothesis test typically is conducted for one postulated truth that is the bottom line to remember.

Now, for the situation where I have at most also I used the equality type. Now at most could be for example, in the automated filling machine example that we looked at in the motivation lecture. We said that we would accept the filling machine or even the manufacturer would consider the filling machine has good, if the variability in the filling across the bottles would be at most 0.01.

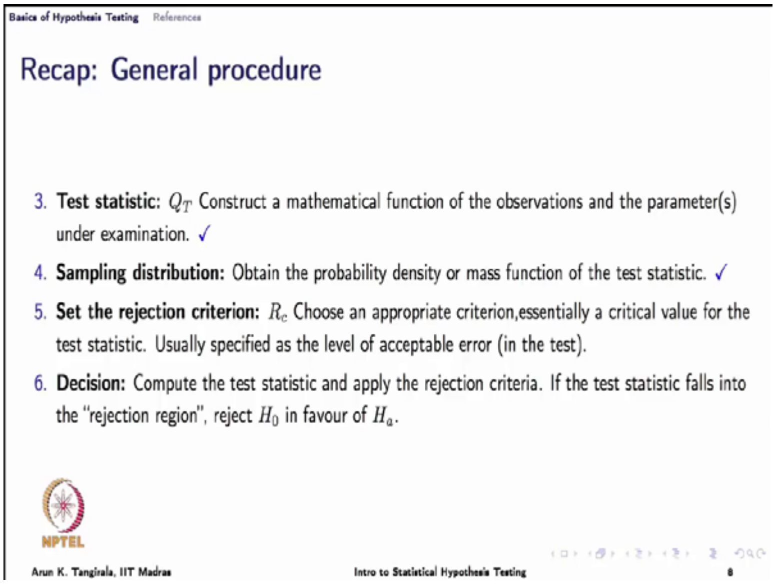
Now, as far as alternate hypothesis is concerned, it is very easy once you stated the null hypothesis the alternate hypothesis is simply the complementary of it so that the null and alternate together span the entire set of possibilities. For instance, if the null hypothesis is truly of the equality type, we are not talking about the practical null hypothesis that is always of the equality type, practically all null hypothesis as we have said is of the equality type. But here we are here saying the problem statement itself involves and equality type hypothesis; in such case, obviously, the alternate hypothesis is of the type of unequal. We would not say inequality, but unequal or that we say that the parameter is not equal to the postulated value; it could be to the left or to the right.

Therefore, we call such a test as two-tailed two-sided test, and I have a small thumbnail of a graphic that we will expand upon later on. I will show you there where you can see 2 small red regions indicated there. Those red regions are nothing but the critical values. And, if the statistic the observed statistics falls into this red region it is kind of an alert region then the null hypothesis is rejected. So, there is an acceptable region and the critical region both computed as probabilities of the statistic taking on certain values and this requires of course, the knowledge the sampling distribution as we have studied in the previous lecture.

Now, of course, what I want to also add is that schematic that I am showing you that the thumbnail that I am showing you is just illustrative it to the curve look symmetric does not mean that the curve which is the sampling distribution of the statistic is always going to be symmetric. It is just to drive home the point of what two-tailed test involves, it involves the testing of statistics to the left and the right of the postulated value and you have critical regions on both sides. Now, when it comes to the other 2 situations, critical regions are either to the left or to the right depending on the whether looking at a lower-tailed or a left tailed test or right or an upper-tailed test. In these 2 situations, the alternate hypothesis are strictly less than and strictly more than naturally, because the null hypothesis of an at least type and the alternate hypothesis should essentially complement the null hypothesis. So, the point is when you are framing the null and alternate hypothesis, first look at the problem statement. And, null hypothesis should always consist of the default; if nothing is stated as a null hypothesis you ask in the absence of data, what would be de facto or status quo or by default kind of statement and then call that as an null hypothesis. And then, whatever you want to test typically goes and sits into the alternate hypothesis.

We will learn through a few examples on how to do that. And again, I want to emphasize what I have written at the bottom of the table. Although, the problem statement may involve null hypothesis of equality, inequality like at least or at most as far as the implementation is concerned with the null and alternate hypothesis framed and of course, the parameter of interest identified.

(Refer Slide Time: 09:41)



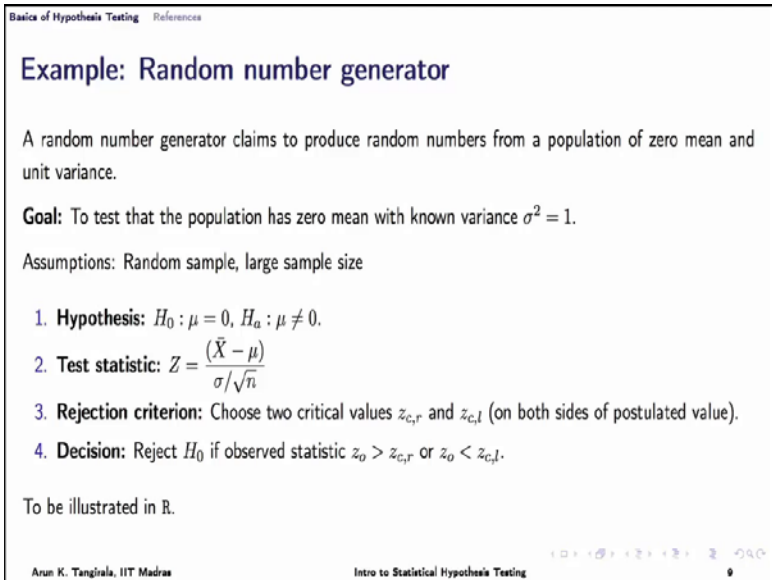
The next step naturally is to choose the test statistic, which we denote by Q T; and test statistic as I have explained earlier can be thought of as a lens that you are an inspector or you are in detective in search of evidence. And, you are using this test statistic as a magnifying lens in search of evidence either to reject h naught most of the times, but if you cannot find enough evidence and then you strict to the null hypothesis. In hypothesis testing as we have learned in the previous lecture, this test statistics nothing but a mathematical function of the observations, it involves the estimate. And, we have also talked about how to choose the statistic or how to choose the estimate that goes into the test statistics; you should satisfy the conditions of unbiasedness, efficiency and consistency sometimes robustness may be required depending on whether the data has outliers or not.

And then, once you have constructed the testing statistic, you obviously need the knowledge of sampling distribution to compute the acceptable and critical regions that is to compute the probability of obtaining statistic as it has been observed whether it falls into the then low probability region and high probability region. For this, we may need the probability density or the probability mass function as the case may be depending on whether the test statistic is a continuous value random variable or a discrete valued random variable. Once we know the sampling distribution then the only step that is left is to set the rejection criteria that is I compute the test statistic and now I want to ask if the null hypothesis is true, do I actually consider this value of test statistic that I have observed acceptable or not. For this I have to set a certain threshold that is what we call as critical value.

Now, this critical value can be specified in different ways; in an elementary way or a rudimentary way, you can directly specify the critical value. When we directly specify the critical value, it does mean something on the type one and type two errors that will shortly essentially the errors that will make in hypothesis testing. On the other hand, in the second approach, we specify the amount of error that we are willing to tolerate in the hypothesis testing from which we derive the critical value. It is a second approach that is widely followed, because then the user knows a priory what is the error that one will incur in the hypothesis test. We saw with the first approach in which the critical value is specified by the user based on some experience or intuition and then the error has to be computed. We will go through both procedures here in this lecture.

Once the critical value is decided using either of the approaches the test statistic is compared with the critical value, and determine whether the test statistic falls in the rejection region or the critical region or the acceptable region and the decision is made. So, this is essentially generic procedure that once sees in a hypothesis test. This is nice because you can now apply this to any hypothesis test problem. What is important is to make sure that at every stage we have made a sound effort to choose for example, the parameter or frame null hypothesis or choosing test statistic and determine the sampling distribution and so on. Once we are sure about our effort and the choices at each stage then the outcome of the hypothesis test is typically reliable one. Now, having said all of that we should keep reminding ourselves that every hypothesis test, no hypothesis test is perfectly let me put it that way; every hypothesis test will incur a some error and that is where we talk of type I and type II errors which we will understand shortly.

(Refer Slide Time: 14:12)



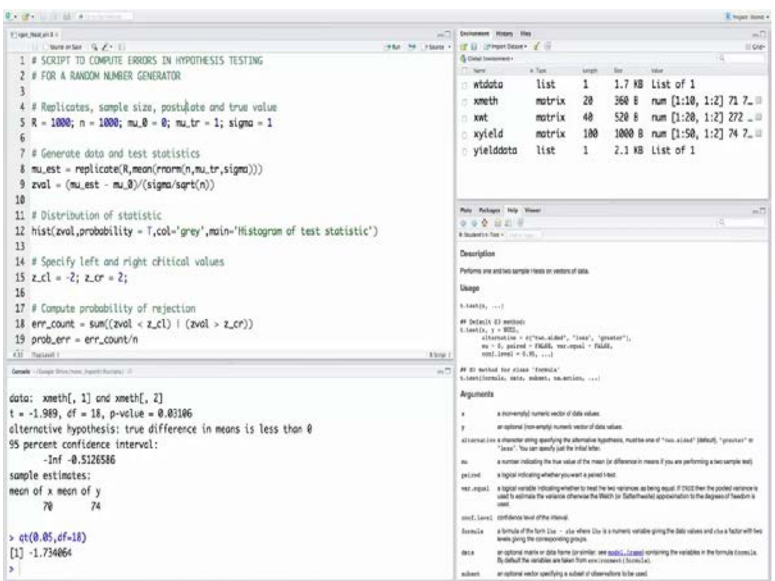
So, let us do that we will define type I, type II error shortly. But before we do that, let us get a feel of what is this type I error through an illustrative example. And in this example, we go back to a our favorite study of random number generator, producing random numbers with 0 mean like the r-norm in R or may be random in mat lab. So, the problem statement is as follows a random number generator claims to produce random numbers from a population of 0 mean; there is no concern about distribution at this moment, from a population of 0 mean and unit variance. Let assume that we have faith in the variance that is variance is unity. So, the only parameter of contention is the mean, therefore the goal is to test the population has 0 mean with the known variance of 1. And, the assumption that we will make is that we will now draw random samples from this population, and the sample size is large.

Now, this large sample size is only because we are going to use the sample mean and we know that the standardized sample mean follows the standard Gaussian distribution regardless of the distribution of the population. If the distribution of the population is Gaussian, then we do not have to make this large sample size assumption. Now, in a step wise fashion, the null hypothesis is obviously, mean is 0 and the alternate hypothesis is that it does not, and the test statistic naturally based on a previous discussions is our standardized sample mean.

And, the rejection criterion now what we can do is we can because it is a two-tailed test you can if you recall from the previous discussion, we do not know if the observed statistic will fall to the right or to the left of the postulated value. Therefore, we need 2 critical values; one for the left and one for the right; we call that as z subscript c l and c r for the left and right ones, respectively. As I said earlier, there are 2 possible ways of specifying those; one it can be a user specified thing. I can specify what the critical values are saying that I know that for this kind of a process, this many samples that probability that is that I will not observe a test statistic as large as z c r or as low as z c l and then go ahead with hypothesis test. Or the second approach, where I specify the error that I will incur and then determine the critical values.

Let us go through the first one, because that approach, because it tells us what is meant by type I error and we will illustrate this in R. Of course, once we decide by the critical values are then the decision is fairly easy to make. Let us now switch out to R may be if you have not opened up R in your laptops, you can actually get fire up R on your laptop and then walk with me through this example, alright. So, what I have done here is an interest of time, I have actually written a script that does this for us. Now, the purpose of this script is not just to perform the hypothesis, but also to illustrate the concept of type one error.

(Refer Slide Time: 17:19)



What is this type of error business before the going to the example? Let me just give you some brief background on it. Remember in hypothesis test, we have a null and we have an alternate, and we make a decision whether the null should be rejected or not rejected. Now obviously, there are different possibilities here; one of the possibilities is that the null hypothesis is true and I end up rejected, right. Indeed for an example, if you manufacture claims that the solid propellant burning rate is on the average 50 centimeter per seconds and I as a quality inspector depending upon how I collected what is the variability and so on end up rejecting it. Or in this example random number generator, let us say that truth is that in the indeed the random number generator, for example, r-norm does give me numbers from Gaussian distribution with 0 mean. And, I end up rejecting that may be because I have such a realization that although I am sampling randomly, there is still a possibility because and choosing a critical of value that can cause this rejection that probability of rejecting the null hypothesis when this true what we call is type I error.

So, let us go now through the example, because I have to calculate the probability that is a chance of making an error, what kind of an error rejecting the null hypothesis. When it is true, I need to actually repeat this experiment of hypothesis testing for different set of samples that is different samples in other words different replicates or realization and then count the number of times I have made a wrong decision. We will assume that it is not true that is indeed r-norm does give me numbers from zero mean. What we shall do is the same a thing that we did in a few other examples previously; we use a replicate command to repeat our experiment of drawing the numbers from Gaussian population.

So, if you look at on the top said the parameters of the simulation r stands for the number of replicates and n, small n stands for the sample size, and then you have mu naught which is a 0. What is mu naught? It is a postulated value and the mu true is the truth value and we said that to 0, because we assume that r-norm in the sense, if I do not specify anything in r-norm you know the default values are 0. There is nothing wrong in setting this is to 1, but then later on I have to take this into account when I count the number of instances I go wrong; for the time being will reset this back to 0. So, the true mean is 0 and true sigma is 1, these are the parameters of the settings of the simulation.

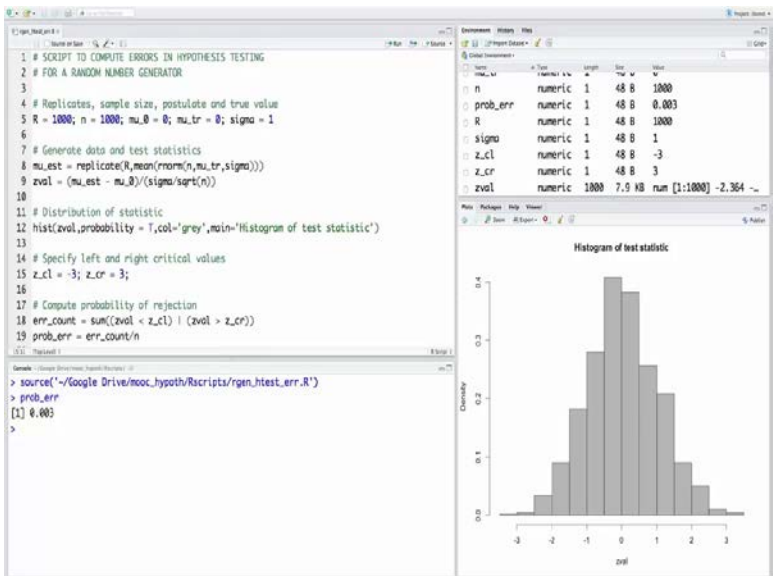
Now, we generate the data and a statistics, and remember we are going to repeat this R that is thousand number of; thousand times and therefore, we put through this mean calculation data generation followed by mean calculation through the replicate. As I have said earlier, you could do this using a for loop, but this is lot more efficient and the commands syntax is fairly obvious in the inner most operation and drawing 1000 samples randomly from a very large population that follows a Gaussian distribution of 0 mean and variance 1. And then, I calculate the sample mean and I repeat this 1000 times and that result is stored in mu underscore e s t. This is the unstandardized sample mean, but our test statistic is a standardized sample mean. So, in the next line, line number 9, I am actually calculating the standardized statistic assuming that I know sigma which any way we have stated earlier.

Now, just as a cross check, we have learned earlier that the sample in follows Gaussian distribution, we have also seen that through the simulation. Just to reconfirm and reinforce this concept we can if we wish plot a histogram of the standardized test statistics and one should expected Gaussian distribution. We will see if that is the case when we run this source code. And finally, now we come to the focus of this script which is to compute the number of instances I make a mistake, when the truth is that mean is 0. For a specify critical value, now let us say because the z that is the standardized test statistics follows a standard Gaussian distribution, we know from Gaussian distribution what kind of extreme values we can expect to see in a standard Gaussian distribution.

And, critical values are something like extreme values, we were saying that the critical value is what you can obtain as an extreme things when the truth when the null hypothesis codes. Of course, a true extreme values are minus infinity and plus infinity, but that is then covering all kinds of situation, no we do not want that because those values can be obtained also when the null hypothesis not true that is a point. We want to ask if the null hypothesis is were true; that means, if the true mean were 0 what are the most likely values that I will get to see for the test statistic. And, based on my understanding of this standard Gaussian distribution, I choose the left and right critical values to be minus 2 and plus 2 that is my choice I can even change into minus 3 and 3 or minus 1 and 1, but because it is Gaussian distribution it is a symmetric distribution the critical regions also have to be symmetric on either side of the postulated value.

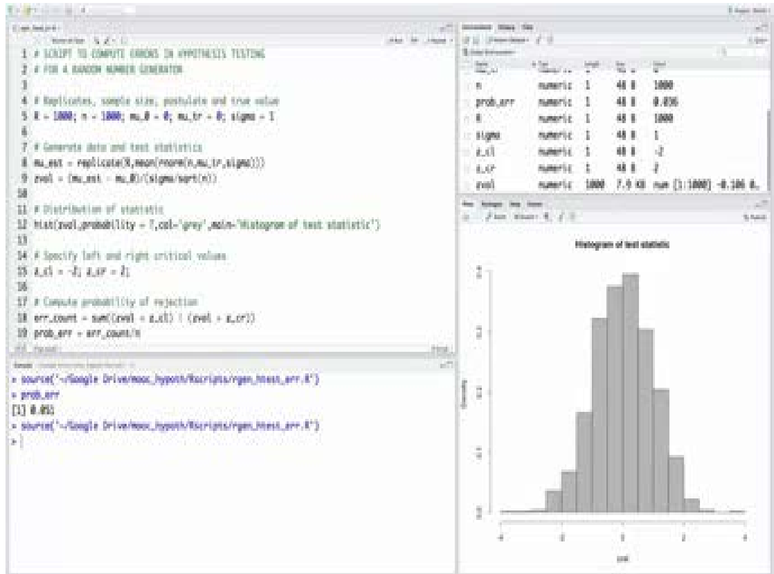
So, let say I choose 2 that is plus or minus 2 as a critical values. Now, in lines 18 and 19, I am counting the number of instances, I would reject the null hypothesis when it is true. What is the criteria for the rejection? If the observed statistics that is a z value is outside this interval of minus 2 and 2, minus 2 and 2 inclusive, then the null hypothesis is rejected that means, I am looking at the instances when the z value is less than to the left that is to the left of the left critical value or more extreme then what I specify as the right critical value. And, the sum here basically it is this id logical operation that I am performing, I am summing up all the number of ones whenever this is case this is the case the result of this logical operation is 1, otherwise it is 0. So, by summing up the result of this logical operation I get a count on the number of instances I have made a mistake and that is reported in some count and probability is essentially the fraction of the instances, that the total number instances is the sample size itself. So, let us run this source code, and see if I can item calculate this let me actually clear all the variables in the work space and I have done that.

(Refer Slide Time: 25:21)



Let me run this, and what I see on the right is a figure which can shows the histogram of the test statistic that I have computed from repeated experiments through the replicate and it shows a Gaussian distribution, it comes as a no surprise, we know this already. And we can see that is centered around mean 0. What is of interest to us is this probability of the error? That is the number of instances that I have made a mistake as fraction of the total sample size. And, you can see that is roughly 0.05.

(Refer Slide Time: 26:03)



In fact, if I run this again, I will get the different property, slightly different property that because you are dealing with different set of 1000 realization and so on. But the point is probability making a mistake, what kind of mistake rejecting it is not when it is true. That is saying that no this random number generator is not giving me numbers from mean 0. Although, what it claims so is very low rejecting that is very low but this is a chance of rejecting that means, you can end up with a realization when R stars are align well, we can end up with a realization which will result in a rejection of the null hypothesis when it is true. This error we call as the type I error and usually this is the denoted by the Greek symbol alpha. And, the significance level is also the name given to this type I error.

What I would encourage you to do is play around with the critical values. For example, we could actually change the values from minus 2 and 2 to minus 3 and 3, and of course make sure you save the file, and before running the source. And of course, before calculate asking whether the type I error will now decrease or increase that is before running the source code, we can (Refer Time: 27:29) and you think of the answer before I run this code for you, whether you expected the type one error to increase or to decrease. Remember, what we are doing by changing from plus or minus 2 to or plus or minus 3, we are extending (Refer Time: 27:45) acceptable (Refer Time: 27:47) we are allowing more and more possible values to be for the test statistic to be acceptable.

So, let us run this now and see what probability we actually get, it is very low that is compared to what we have seen for minus 2 and 2. In fact, now you should expect that as you extend the acceptable region, the type I errors start shrinking. Ideally, you may think, I mean we may think that why not actually set the acceptable region to may be minus 10 plus or minus 10 or even plus or minus 100, so that I do not make any errors at all in rejecting the null hypothesis. Well, life is not so easy, once you try to minimize this type I error, there is some other error which will shoot up and that is what we call as a type II error. And, let us now go to the definition and see, why this is so? What happens when we try to reduce a type I error? Where else are we really affecting the problem?