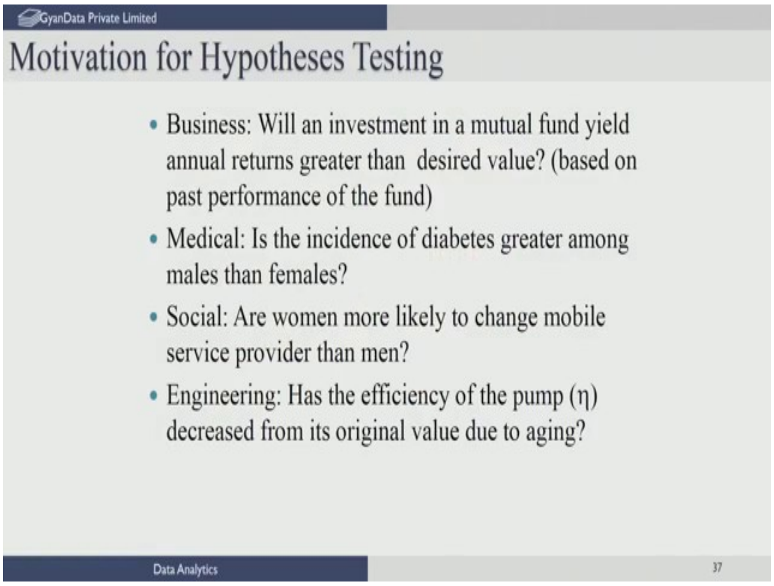
Data Science for Engineers
Prof. Raghunathan Rengaswamy
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Lecture – 22
Hypotheses Testing

Welcome to this last lecture on Introduction to Probability and Statistics. In this lecture we will introduce you to the basics of hypothesis testing which is an important activity when you want to make decision from a set of data.

(Refer Slide Time: 00:39)

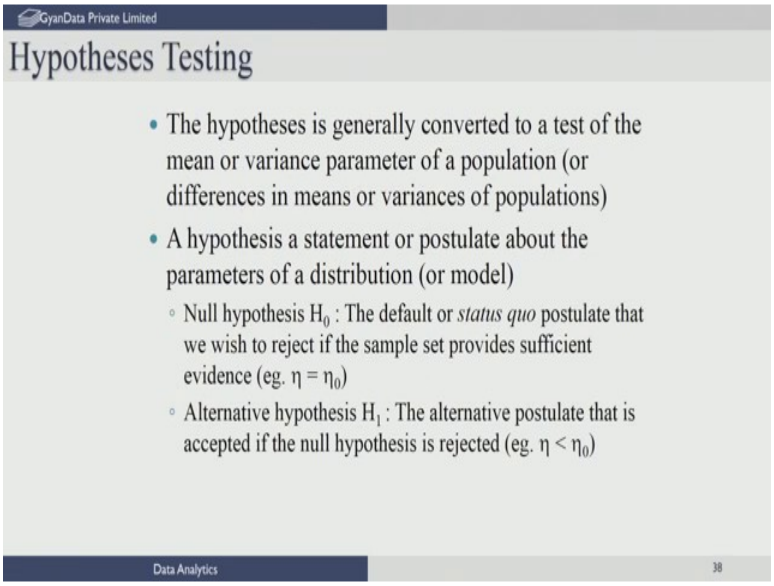


So, to give you some motivation for hypothesis testing we look at some cases. Let us look at a business case where you are interested in investing in a mutual fund and you want to know whether this investment will yield a certain annual return greater than some desired value let us say 15 percent or 20 percent that you might want. Now this decision has to be made based on past performance of the fund which you have data which you can collect.

Similarly, let us assume you are a medical practitioner and you want to ask this question whether the incidence of diabetes is greater among males than females based on data that you have gathered about males and females and what proportion of males and what proportion of females have diabetes. Now, you can ask similar question in the social sector. You can if you are a service provider mobile service provider you want to know whether women are more likely to change service provider than men and depending on that you might want to provide more incentives accordingly for women to retain them.

In engineering you can ask a question such as a pump efficiency whether it has degraded from its original value due to aging and if so you may want to do some maintenance of the pump. So, these kind of questions that are decisions that you have to make will be based on data that you have gathered about the particular object or item.

(Refer Slide Time: 02:02)



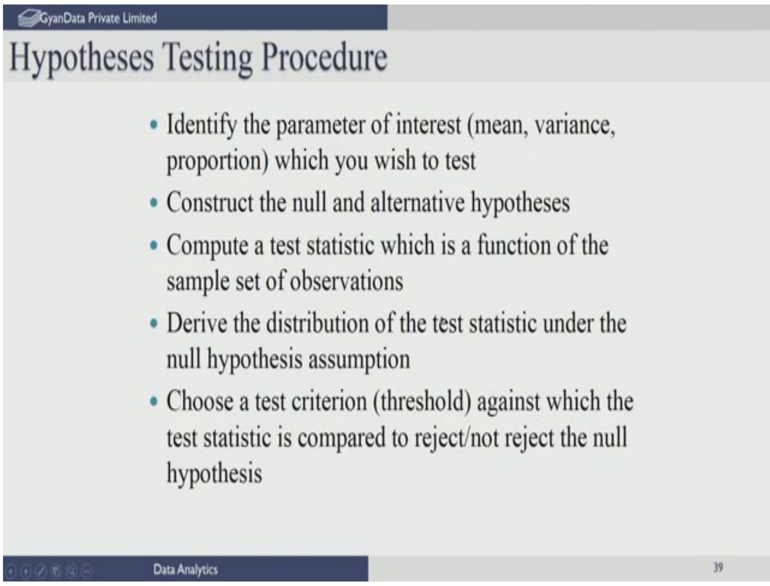
Now, let us look at how do you perform these decision making using what is called hypothesis testing. Typically what you have to do is to convert this hypothesis into a test for the mean or variance parameter of a population or perhaps a difference in the means of two populations or the difference of variances of the two population.

The statements that previously we saw have to be first converted into a test of a parameter of some population. Now, this hypothesis is a postulate about the parameters. You call them either the null hypothesis or the alternate hypothesis. The null hypothesis, the default hypothesis, that you want to test, the status quo postulate that you wish to reject if the sample set provides sufficient evidence. For example, this is the statement about which you want to make the strongest claim based on the data. So, that you choose as the null hypothesis and what we call the default hypothesis. Example you want you want to convert this statement into a parameter called η, whether this η efficiency, let us say of the pump = η_0, its original value.

If you have per performance data that you collect then you based on the data you want to reject whether this efficiency at current time is different from its original value η_0. If there is evidence, you will reject this hypothesis in favor of the alternative hypothesis which is essentially saying that the pump efficiency in this case is less than η_0. So, this is called the alternate hypothesis. So, you set up the hypothesis such as $\eta = \eta_0$ which you call the null hypothesis and the alternative hypothesis which you want to choose in favor of this if the evidence is there from the data such as for example, η less than η_0.

So, all hypothesis testing has this null and alternative. And we can have different types depending on whether you are testing for the mean or the variance and whether this is less than or greater than or on both sided and we would see many such examples.

(Refer Slide Time: 04:20)

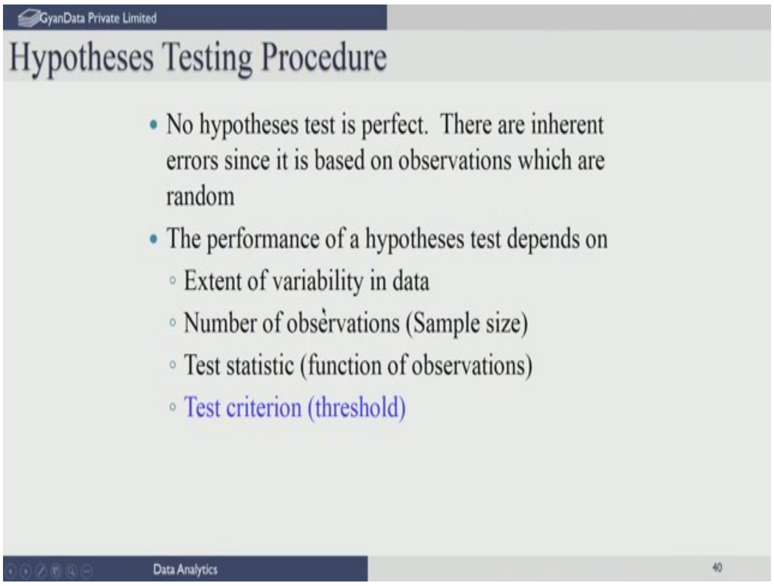


Now, in order to perform this hypothesis testing and make a decision; As we said the first step is to identify the parameter of interest which you wish to test, it could be the mean, it could be the variance of the population or the proportion of the population that you want to verify value. Now, you construct the null and alternative hypothesis as we said before. Then, based on a data experimental data about the system you collect and then you construct something called the test statistic. This is a function of the observations.

Now, this could be for example, if you are testing for the population mean you may use as the test statistic, the sample mean. If you are testing for the population variance you may use as test statistic the sample variance and so on, so forth. Now, you also have to derive the distribution of the test statistic under the null hypothesis, what it means is if the null hypothesis is true what is the distribution of this test statistic that you are computed?

Now, based on this distribution you choose a threshold or the test criterion and now you compare the computed test statistic against this criterion and decide whether to reject the null hypothesis or not reject the null hypothesis. This is the overall procedure. We will show through examples how these procedure is carried out for different cases.

(Refer Slide Time: 05:46)

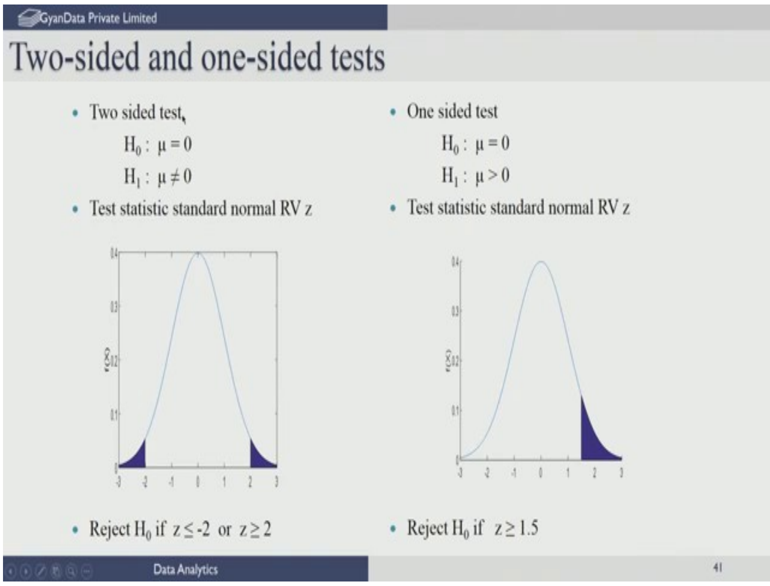


Now, you have to understand that the hypothesis testing procedure is an imperfect decision making process. Why? Because you are basing the test on a sample set of observations not on the entire population. So, there are inherent observe errors in the hypothesis testing because it is based on observations which are themselves stochastic in nature. Now, therefore, the performance of the hypothesis test, how well this decision making perform, depends on how much variability is there in the data. This you may not be able to do much about this they might be enough sufficient variability in the data which prevents you from making a good decision.

You can also alter the performance of the test by choosing the number of observations, experimental observations, you want to make which is called the sample size. Now, the test statistic as we said is a function of the observations there are different sometimes you might have different choices of functions, and some test statistic may actually perform better than others. That depends on the theoretical foundations of statistical hypothesis testing, we will not touch upon this. We have control over the of number of observations. We will take a look at how we can alter the performance based on sample size.

Finally, you also choose a threshold against which you are comparing the test statistic and therefore, you can alter the performance based on the criterion that you select, and we will see how this test criterion affects the performance of the hypothesis test.

(Refer Slide Time: 07:17)



Now, there are two types of hypothesis tests what we call the two sided test and the one sided test and I am giving a simple illustration to tell you what a two sided test means. Suppose you are testing for the population parameter, population mean μ, and you want to test the test whether this population parameter = 0 or != 0. So, the null hypothesis is the population parameter is 0, and the alternative is then a population mean is != 0. So, you observe some set of observations from this particular population we have a sample and you have computed let us say a sample statistic and let us also assume that sample statistic happens to be a standard normal variable z. We will show you how to construct a test statistic that finally, has this kind of a distribution for testing the mean.

But let us for the time being take it that we have a test statistic based on the observations we have made and this test statistic that we have computed is the standard normal test statistic z. Now, we know that this z, because it is a standard normal distribution, will have some shape like this and about 95 per-cent of the time the data the statistic will have a value between around - 2 to + 2. So, for a two sided test, we will say if the test statistic happens to be very large we will reject it or if it is very small we will reject it. Why? because if it is very large it means it does not come from a distribution with mean 0. So, in this particular case what we mean by very large we can choose a threshold let us say - 2, I am sorry + 2 and if the statistic is greater than 2 we reject the null hypothesis or if its small what do we mean by small if it is less than - 2 we reject the null hypothesis.

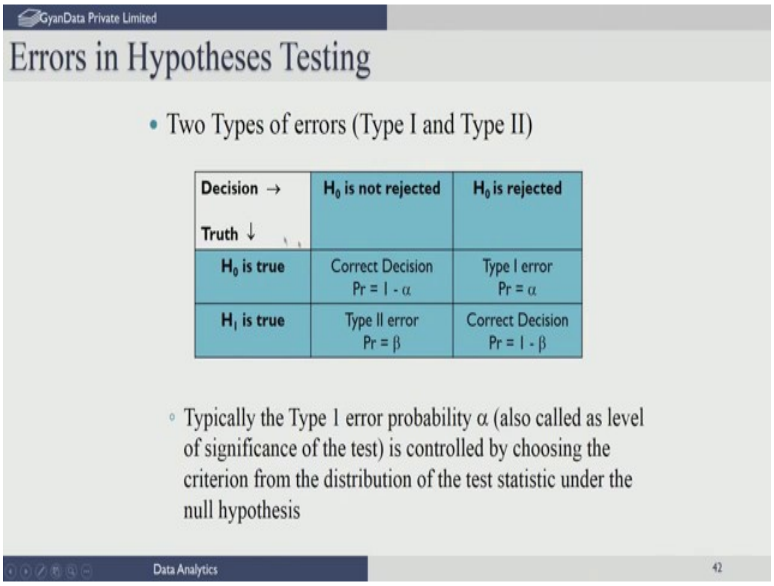
So, in this case we reject the null hypothesis whether the test statistic is less than a particular value the threshold value, in this case I have chosen the threshold as - 2, or if the statistic is greater than the threshold value, which is + 2. So, there are 2 thresholds the upper threshold which is 2 and the lower threshold which is - 2 because it is a two sided test. We want to reject the null hypothesis if μ is less than 0 or μ is greater than 0. How we choose these thresholds and what are the implications we will see later. But realize that if it is a two sided test you basically have a lower criterion threshold and the upper threshold which you select from the appropriate distribution.

Now, suppose you have the same thing, but you are only interested in testing whether the mean is 0 or greater than 0. So, in this case the null hypothesis is μ = 0 and the alternative is μ less than 0 or greater than 0. So, notice that μ = 0 implies that you are not interested in the case when μ is less than 0, you are you are not going to reject the null hypothesis if μ is less than 0 you are going to reject the null hypothesis only μ is greater than 0. That is why you have written the alternative like this. This is called the one sided test.

In this case let us assume that you again have computed a test statistic based on the observations and that is a standard normal test , then we only have an upper threshold, because we want to reject the null hypothesis only if the mean is greater than 0. So, similarly if the statistic is greater than a threshold then we reject the null hypothesis. So, we have a upper threshold, in this case I have chosen 1.5 and reject the null hypothesis if the statistic happens to be greater than 1.5. If it has a low value we are not bothered because we are not bothered when μ is less than 0.

So, although technically we should call the null hypothesis μ less than or = 0 it is usually stay in equality as μ = 0 with indifference. Essentially the alternative tells you whether we are indifferent to μ less than 0 or not. So, this is called a one sided test. So, depending on the type of test whether its two sided or one sided you choose thresholds and then compare the test statistics against those thresholds.

(Refer Slide Time: 11:41)



Now, when you do such a test you commit two types of errors. So, essentially let us look at this truth table. Suppose the null hypothesis is actually true and you have made a decision to not reject the null hypothesis which means you have made the correct decision. So, this will not happen all the time because your sample is random it is possible that even if you are not high passes is true, you may conclude that the null hypothesis you may decide to reject the null hypothesis in which you commit a type I error what we call a type I error or a false alarm.

So, when the null hypothesis is true and then you reject the null hypothesis based on the sample data and your statistic and the threshold d of selection. So, you have selected, we call this a type I error or false alarm and the probability of that is known as α and we call it a type I error probability.

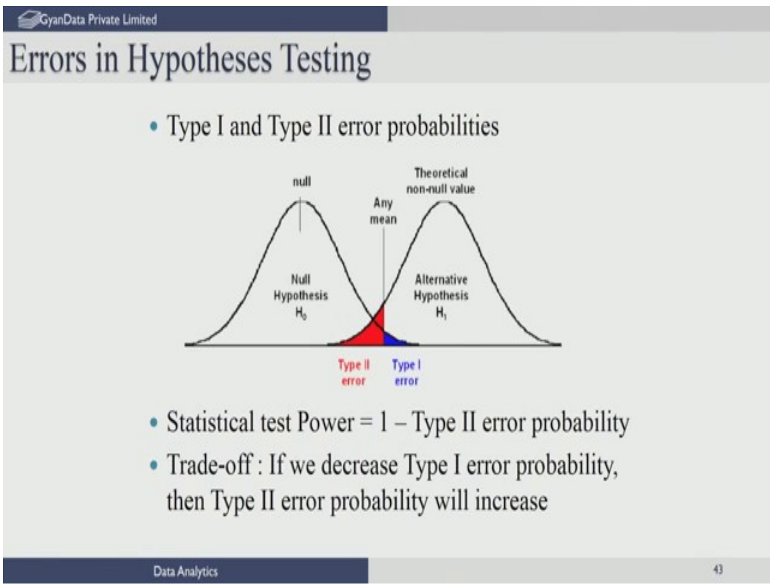
So, you will have it is not that your decision is perfect you will always commit some type I error α depending on the threshold that you have selected and the statistic you have computed. Similarly let us assume that the truth is that the alternative is correct. So, in this case it may turn out that from your sample setup data you do not reject the null hypothesis, in which case you commit what we call a type II error and this type II error also has a probability which is denoted by β.

On the other hand if the alternative hypothesis is true and do you do reject the null hypothesis in favor of the alternative then you have made a correct decision and that correct decision probability is known as power of the statistical test and is denoted by 1 - β. Remember the only one of two decisions you have always going to make, you are either going to reject the null hypothesis or you are not going to reject it. So, the total property always be 1 and the probability of type II error if its β then the probability or statistical power is 1 - β. So, there are two types of errors that you commit what to call the type I error probability α, and the type II error probability β.

So, the type I error probability α is also known as the level of significance of the test and this is typically what you control. You do not control the type II error probability. When you construct this hypothesis test to construct the test statistic and choose a threshold you basically try to control this type I error probability. The type II error probability results as a consequence of this. So, we will choose the criteria from the distribution of the test statistic under the null hypothesis because the null hypothesis is very precisely stated. If you go back and look at it we are able to state the null hypothesis precisely, the parameter value is 0.

Whereas, the alternative hypothesis we do not know what the parameter value we entertain a huge set $\mu_0 = 0$ involves everything other than 0. So, the parameter set is undefined very not clearly defined at least exact value is not clearly defined in the alternative. Whereas, for the null we are clearly defined in both cases. Therefore, it is possible to construct the distribution of the test statistic under the null hypothesis and that is the reason we only end up con controlling the type I error probability and not the type II directly by choosing the test statistic.

(Refer Slide Time: 15:13)



Now, let us see how we actually control the type I error probability by choosing the appropriate test statistic value, test criterion value. So, let us look at the qualitative comparison of the type and type I error and type II error.

Let us assume that under the null hypothesis you have this distribution which is known as distribution of the test statistic under h naught. Now, depending on, remember that even if h naught is true, the statistic can have a value between - ∞ to + ∞. On the other hand you are going to choose some threshold in I have indicated this by this vertical line you are going to choose some threshold because you cannot say that you will accept you will not reject the null hypothesis whether the value is minor any value between - ∞ and + ∞. In which case you will never reject a null hypothesis whatever be the evidence you get that is not fair.

You decide to null reject the null hypothesis let us say in this case a one sided test and you have decided to reject it, if the statistic exceeds this threshold value. In which case notice that when the null hypothesis is true this test statistic can have a value greater than this threshold. And the probability that can have a value greater than the threshold is indicated by this blue area. So, this is your type I error probability that you have committed . If the null hypothesis is true your test statistic can exceed this threshold in which case this is the area the probability that that the test statistic can exceed this threshold and therefore, this is your type I error probability.

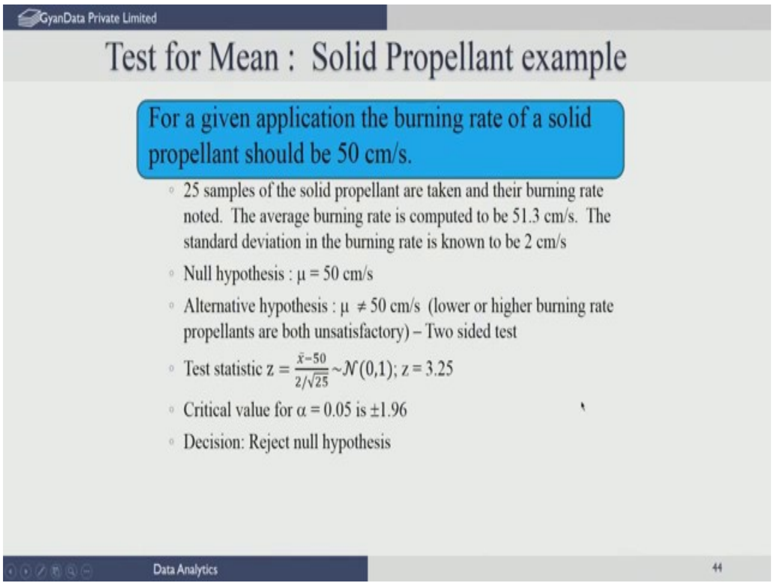
Now, if you move the threshold to the right obviously, you can reduce your type I error probability, but there is a price to be paid. Let us look at what is the price. Let us look at what happens to a type II error probability. So, let us assume that the actual distribution of the test statistic under h 1 under the alternative happens to be this kind of a distribution. Of course, this depends on what value the parameter is going to take.

Suppose we consider this distribution as the distribution of the test statistic if the alternative is true, then what happens is there is a probability that the test statistic will take a value less than this threshold and that is given by the probability marked in red, which means that even if the alternative is true you will not reject the null hypothesis because your test statistic is falling to the left of this threshold. And the probability that the statistic will be left of the threshold is given by the red area and that is going to be of type II error prob-ability. And the power is just 1 - this which is the area to the right of this under this distribution, under h 1. Now, notice that as you move the threshold to the right, the blue area shrinks which means your type I error shrinks, but your red area increases which means your type II error increases. So, there is a trade off.

If you try to make your test perfect in the sense of no type I error then you will come commit a type II error of one. Which means you are be a very insensitive test you will never be able to find whether your mean is different from 0 for example. So, the more less type I error you are willing to entertain the less sensitive your test will be. So, that is always a trade off you cannot help it. So, that there is a trade off if we decrease type I error probability then type II error probability will increase there is no choice and you have to accept this trade off .

So, you decide that I am willing to accept a type I error probability of 1 per-cent or 5 percent or 10 percent, and accordingly your test will be less or more sensitive. This is the way, the threshold selection is based on how much of false alarm probability or type I error probability you are willing to accept and that is the trade off you should accept and accordingly your test will be less or more sensitive.

(Refer Slide Time: 19:16)



Now, let us look at some examples for hypothesis testing. In this case we have looked at a manufacturer of a solid propellant and we want this solid propellant to burn at a certain rate and this burning rate of the solid propellant is specified to be let us say 50 cm per second. If it burns at a higher rate then we will not be able to come control the rocket, if it burns at the slower rate then the rocket may not even take off . So, we are going to check whether the propellant we are made, which is based on mixing a lot of different chemicals, whether it will have a burning rate of 50 centimeter per second.

Now, what we have done in the from the mixing bowl where we are making the solid propellant, we have taken 25 samples from different locations in the mixing bowl and each of these samples we test in the lab and find that what their burning rate is. And it will be some value maybe it is 48, 49, 51 whatever. We compute the average of these 25 samples.

Notice, that the population parameter mean that we are testing is μ = 50 and we are going to use the sample mean. We have taken 25 samples we compute them average or the sample mean of the burning rates of these 25 samples and we are going to make a judgment about the entire batch that we are making in the mixture remember. So, the population in this case is the entire batch of product that you are making and you are taking only a few samples and based on these samples you are making a judgment about the entire batch. So, the population parameter mean μ is what is what you are interested in you do not know what this value will be, you are going to ask this question whether that mean is going to be 50.

The sample mean that you have computed based on these 25 samples happens to be 51.3 centimeter per second, the burning rate average value. We have also computed the standard deviation based on these 25 samples we can compute the standard deviation of the sample and we find that the standard deviation is two centimeter per second. So, different samples have different burning rates and we find that the variability or the standard deviation in the burning rate of the samples is 2 centimeter per second.

Now, based on this data that we are collected the sample mean and the sample standard deviation we want to ask the question whether the population mean happens to be 50 centimeter per second. I am sorry let us say that the sample standard deviation is already known to you given to you that it will be 20, 2 cm per second it is not based on the sample its already known to you. Let us take that as a case which is the simplest case.

So, here we have actually said the null hypothesis population as having a burning rate of 50 centimeter per second and the alternative is μ_0 = 50. Notice, it is a two sided test because we want to reject this batch if the burning rate is less than 50 centimeter per second it is not useful to us or if its greater than 50 centimeter per second that is why we have taken the alternative to be not = 50 centimeter per second. Now, this is a two sided test. Now,how do we construct the tests? I already talked to you that the sample mean has a normal distribution with population mean as the parameter the expected value and the standard deviation of the population divided by square root of n.

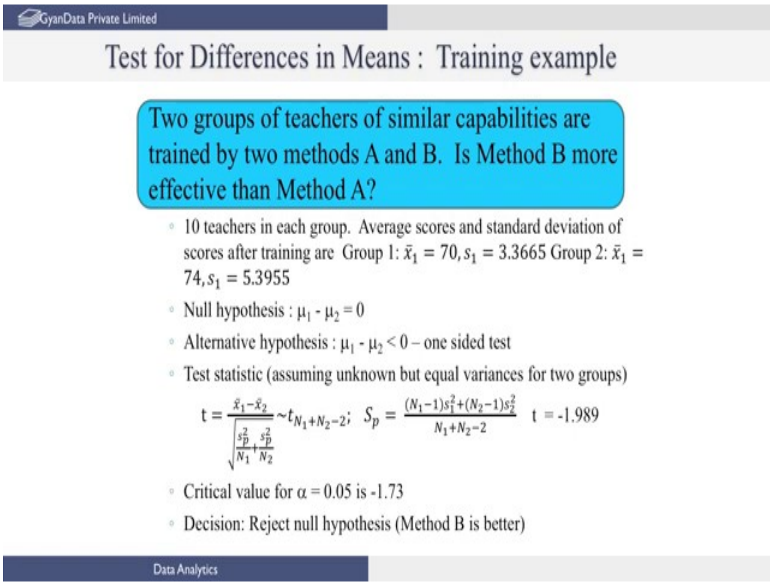
You can refer to the previous lecture to find out that the distribution of $\bar{x}$ the sample mean, is the same as the population mean and has a standard deviation which is one standard deviation of the population divided by the square root of the number of samples you have taken. So, we can now do what is called standardization which is subtracting the mean of the sample mean which is 50, the expected value ,divided by the standard deviation of the sample mean, which is 2 by root 25, and we get what is called the standardized value and this standardized value z will have a standard normal distribution with 0 mean and unit variance.

Now, notice that this will be the distribution if the population mean happens to be 50; that means, under h naught under the null hypothesis this test statistic will have a standard normal distribution. Now, we know that 95 percent of the of a standard normal variable will lie between + or 1 - 2 standard deviation approximately. More particularly the standard normal variable will lie between + or - 1.96, 95 percent of the time. So, if we are willing to tolerate a type I error probability of 0.5 percent let us say, which means what you are saying is, the area that you have on the left + the area that you have on the right of the upper threshold = 5 percent. If you are willing to tolerate that type I error probability then you can choose your criterion or your I am sorry your threshold as 1.96 and - 1.96 and that is exactly what is stated here.

The threshold value is + or - 1.96 and if your z statistic happens to be greater than 1.96 you reject it if, if it is less than one point - 1.96 you reject the null hypothesis in this particular case. If you substitute the value of the $\overline{x}$ that we have obtained which is 51.3 and we have this standard deviation which you already know and we compute this we get a value of 3.25 which is greater than 1.96 and therefore, we reject the null hypothesis - this batch does not satisfy our needs. So, this is the way of testing and this is an example of testing for the mean given the standard deviation of the population.

In case your standard deviation of the population is not known, then you use the sample standard deviation and this test statistic we can show is a t statistic and therefore, you can derive your thresholds from your t distribution rather than the z distribution.

(Refer Slide Time: 25:41)



Let us look at another example where we are testing for difference in means. Here we have two groups of teachers. This is a social example. So, we have two groups of teachers of let us say similar capabilities. There are 10 teachers in each group and we have trained them by two different methods A and B teaching methodology.

Now, we want to know whether method B is more effective than method A. Now, let us actually we assumed that we have 10 teachers in each group you have given training for one group of teachers using method A and another group of teachers using method B. And then we test them for the effectiveness and we find this course, we find that the average course of group 1 happens to be 70 with a standard deviation in the marks is 3.37, and group 2 has a sample mean of 74 with a standard deviation as 5.4 approximately.

So, we want to know whether method A and method B are both equally effective or method B is more effective than method A. This is the question that we want to do. So, we have chosen the null hypothesis that method A mean and method B mean both are equal which means their difference in the mean should be equal to 0 that is your null hypothesis. We will reject this in favor of the alternative which says the average the effectiveness of method A is less than the effectiveness of method B which means the mean of method A is less than the mean of method B , which means the difference should be less than 0. Notice that it is a one sided test because we are not interested in looking at whether method A is more effective than method B, we are only asking the question is method B more effective than method A and we are setting up our hypotheses accordingly. The difference in means is equal to 0 happens to be the null hypothesis, the difference in means 1 - 2 is less than 0 is the alternative we choose, in case we find enough evidence for rejecting the null hypothesis.

Now, we will assume that the standard deviation of these two groups is same, although we do not know the standard deviations. So, we will take make the assumption that the standard deviation of marks obtained by group A population and the standard deviation of the group B population of teachers are equal, but unknown. But, however, we have the sample standard deviations from these two groups from which we can estimate the standard deviation of this entire group of teachers.

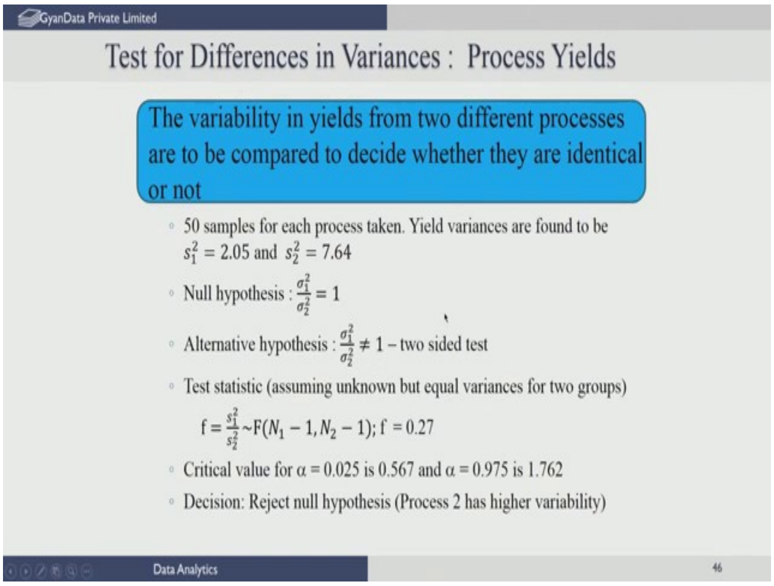
Notice that all the teachers I can group them because they have the same variability, there is no difference in the variability of group A and group B. So, we can pool their variances and we can obtain a pooled variance by just taking the sum square deviation of all with their respective means and then obtaining a pooled variance. This is a way by which you obtain the pooled variance for two groups, and once you obtain an estimate of the standard deviation of the group of teachers then you can take the difference in the sample means because remember you are testing the difference in means, so you can take as the test statistic in the sample means ($\overline{x}_1$ - $\overline{x}_{2}$) divided by the variance standard deviation of these means which is what this is all about. So, S p represents the standard deviation of the difference in the means, remember assuming that we have they are both the groups have the same variances.

So, now, we can show that this particular statistic we have computed is t statistic because σ is also estimated from the data. So, we can now compare this with this t distribution, the number of degrees of the t distribution happens to be N1 + N2 - 2. Notice that that depends on the denominator degrees of freedom which is essentially total number of observations - 2 parameters which are mean parameters that you have used up to estimate the means. So, the remaining degrees of freedom is this and that is how this number of number of degrees of freedom comes about.

In fact, for large enough N1 and N2, if you take large in a frame, you can actually approximate this with a standard normal variable , but in this case let us let us do a precise job. We will choose the test statistic test criteria from the t distribution with this many degrees of freedom 10 + 10 - 2 which is 18 degrees of freedom and we find that the one sided confidence interval. Remember this is one sided if we are willing to tolerate a type I error probability of 5 percent then there is only a lower threshold less than 0 notice. So, - 1.73 the probability that a t distribution with 18 degrees of freedom is greater than this - 1.73 will be 5, 5 percent.

So, if we are willing to tolerate the type I error probability of 5 percent then we can choose the threshold value as - 1.73 drawn from the t distribution with eighteen degrees of freedom and compare it with the statistic. The test statistic itself we compute by plugging in the value for $\overline{x}_1$, $\overline{x}_2$ and so on so forth we get - 1.989, since this is less than - 1.73 we reject this null hypothesis in favor of this alternative, which means method B is more effective that method A that is a conclusion.

(Refer Slide Time: 31:27)



So, let us look at another example which is a test for difference in variances. Now, we have two processes which have different variability and we want to compare whether these two process have the same variability or not. So, we have taken 50 samples from each process and computed their variances and found that one has a variability of 2.05 and the other has a variability or a variance of 7.64. Now, what we want to compare is whether these two variances, population variances, of these two process are equal or not which means the ratio of the variances we want to check whether it = 1 or not.

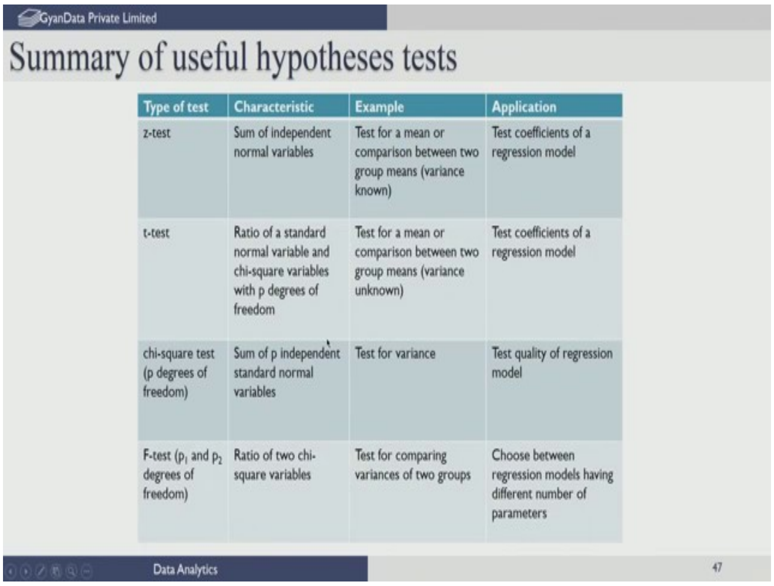
The alternative hypothesis is that these variances are different, so we are asking whether the ratio of the variance is not equal to 1. It could be that $\sigma 1^2$ is less than $\sigma\ ^2$ or $\sigma 1^2$ is greater than $\sigma 2^2$ both cases we want to reject the null hypothesis. So, this is a two sided test.

Again we can use the ratio of the sample variances as a test statistic and this ratio turns out to be an f distribution with degrees of freedom N1 - 1 and N2 - 1 where N1 and N2 represents a number of samples we have taken for each of the process. In this case we have equal number of samples we have taken, even if you are taken unequal samples we can appropriately choose the degrees of freedom for the f distribution and compare. So, in this case if you plug in the values for S 1 squared , S 2 squared we have got a f value of 0.27, and if you go to the f distribution with N1 - 1 which is 49 degrees of freedom and N2 - 1 which is 49 degrees of freedom and ask 5 percent probability which we break it up as two, left of the threshold should be 2.5 percent and greater than the threshold is 2.5 percent.

So, we ask what is the threshold where lower threshold value for the f distribution and it turns out to be 0.567 that means 2.5 percent, there is a 2.5 percent probability that this f distribution is less than this value. Similarly under h naught and similarly there is A_2.5 percent probability that the f distribution is greater than this value. So, the lower threshold we choose as 0.567 upper threshold is 1.762.

The statistic we are computed happens to be less than this. So, we reject the null hypothesis in favor of the alternative and claim that that the two variances of the two process are not equal. Of course, although this implies that the process two has a higher variability, if you really wanted to test this and not worry about process one then we had have set it up differently, the hypothesis testing.

(Refer Slide Time: 34:24)



In summary I want to just go over some of the standard tests that we use in hypothesis testing and see where they are useful. So, the z-test or the standard normal variable statistic which has a standard normal variable, we call it a z-test. And this is typically used for testing of the mean or a comparison between two means when the variance of the population is known.

The application of such a test is usually found in testing whether the coefficient of a regression linear regression model is 0 or not. Now, the t test on the other hand is used also for a test of the mean or a comparison between means group means, but in this case the population variance is not known. So, we use the sample variance for normalization and that gives rise to a t test. Here also we whenever we test for the coefficients of the regression model under the assumption that the errors corrupting the variables, we do not have an idea the way a standard deviation of the errors that corrupt the observations are unknown, then we use the t test and of to test whether the coefficient of regression model = 0 or not.

The χ square test is used for testing the variance of a sample and the variance of a sample, this test is used to whether a regression model is high is good or not, whether acceptable quality or not. The objective function of a regression model is the sum squared term and is similar to a variance, and you can use it to this χ square test to test whether the integration model is acceptable or not.

The f test is used when for comparing variances of two groups. In the case of linear regression this is used to choose between two regression models having different number of parameter. Sometimes you might actually build a model by dropping a variable and you want to know whether the model you have built by reducing the number of variables is better than the 1, that by retaining all of them, in such a case this f test comes in very handy.

We have completed the introduction to probability and statistics. We will see you in the next lecture which we will talk about linear regression and the application of these concepts to linear regression.

Thank you.