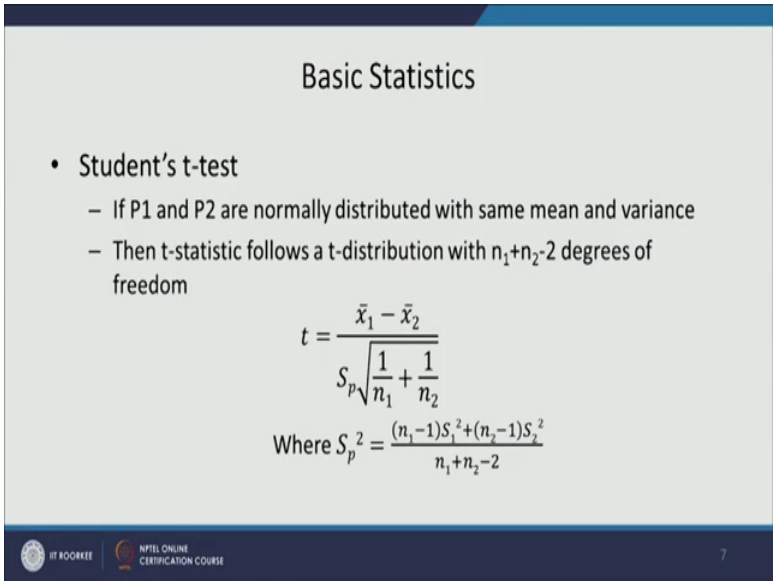
Business Analytics & Data Mining Modeling Using R
Dr. Gaurav Dixit
Department of Management Studies
Indian Institute of Technology Roorkee

Lecture - 05
Basic Statistics- Part II

Welcome to the course business analytics and data mining modelling using R. This is a 4th supplementary lecture on basic statistics. So, in the previous lecture we stopped at; we stopped our discussion about student's t test, so let us pick up from there. So, as we discussed in the previous lecture the most common type of testing that we actual do hypothesis testing that we actually do is about difference of means.

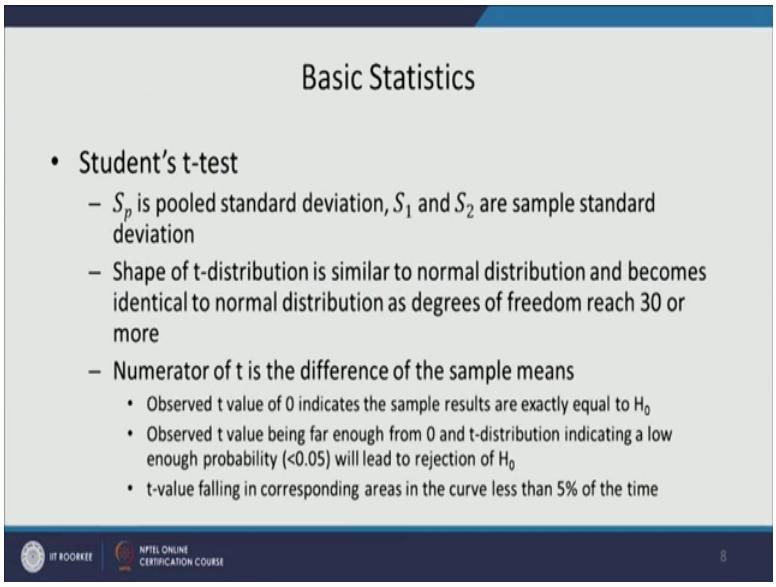
(Refer Slide Time: 00:56)



So, let us take this example if P1 and P2 are normally distributed with same mean and variance, then t-statistic follow a t-distribution with n1 plus n2 minus 2 degrees of freedom. So, we talked about this particular formula of t-statistic in the previous lecture as well. So, you can see t can be computed as difference between x1 bar and x2 bar divided by pooled sample variance and then multiplied by this factor in the square root; 1/n1 + 1/ n2. How the pooled sample variance is computed? Can also you can also see here, it is kind of a weighted average of sample variances from our population 1 and population 2, you can see S p square is (n1-1) is (n1-1) $S1^2$ + (n2 –1) $S2^2$/ n1+ n2- 2.

Now, as far as t-distribution is concerned; shape of t-distribution is concerned it is quite similar to normal distribution and as the number of observation; degrees of freedom these 30 or more, it closely resembles to normal distribution.

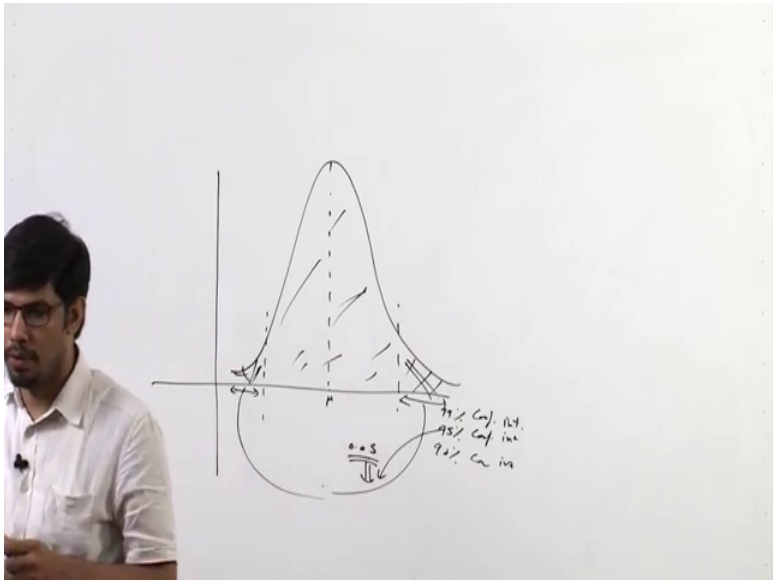
(Refer Slide Time: 02:10)



So, both t-distribution and normal distribution they are bell shaped curves. Now, some specific discussion points on the t-statistic formula; you can see the numerator of t-statistic is actually the difference of sample means. So, from there you can understand for our null hypothesis and alternate hypothesis for their testing, if the observed t value is actually is comes out be 0, then that would actually indicate that sample results exactly equal to null hypothesis, that means, the means are equal.

Similarly, if the observed t value, that is far enough from 0 and t distribution indicates a low enough probability let us say less than 0.05, then in that case our null hypothesis H0 would actually be rejected. So, to get a, better understanding let us look at the normal curve.

(Refer Slide Time: 03:18)



So, let us say, so for a value, a t value which is far enough from 0 and the probability of that particular t value falling within this curve is quite low, let us say 0.05 or less, then in that case null hypothesis would be rejected because our t value is falling in these 2, one of these 2 areas. So, our t value is not falling within this main area and probability of falling in this particular area is low which is 0.05, that means, there are more chances that particular sample results; that particular sample is falling in these 2 smaller regions.

So, therefore, in this case normal in this case, a null hypothesis would actually be rejected. The same discussion can also help us in understanding the confidence interval which we will cover later on in this lecture, this is about you might have come across the words like 95 percent confidence interval, 90 percent confidence interval, 99 percent confidence interval. So, these words terms you might come across, so in this case though we have taken this we talked about a small probability and the t-statistics falling in this particular region, this is corresponding to 95 percent confidence interval. So, there more discussion on confidence interval we will do later on this, later during this lecture.

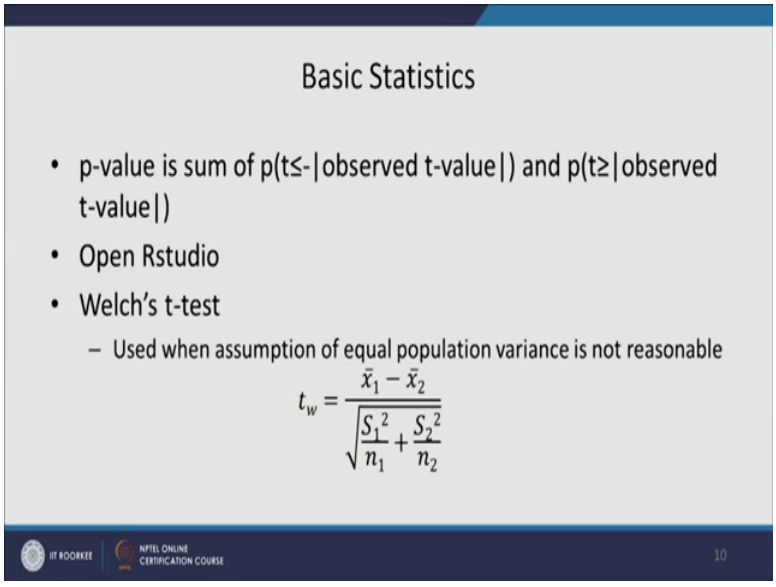
So, another way to understand this is the t value falling in the corresponding areas in the curve is less than 5 percent of the time, so that is another way of understanding this. So, the low probability that we talked about 0.05, it is generally denoted using alpha and this also known as significance level of the test.

Now, how do we find out whether a null hypothesis is going to be rejected or not, so we generally compute this t asterisk value, which is determined in such a way, that probability of magnitude of observed t value being greater than this t asterisk value is actually alpha. So, in such a fashion t asterisk is determined for different values of observed t's and once that t asterisk value is determined, we generally compare it using the observed t value and if the magnitude of the observed t value is greater than t asterisk, then a null hypothesis is actually rejected.

You can see t asterisk is determined such that P and probability of absolute t greater than or equal to t asterisk is alpha which is 0.05, for an example and third point is about that null hypothesis is rejected, if observed value of t is such that absolute value of t is greater than or equal to t asterisk.

Now, significance level of statistical test is the probability of rejecting the null hypothesis for example, in this particular case we assume that alpha is 0.05. So, if null hypothesis is true and alpha is 0.05, then the observed magnitude of t would actually exceed t asterisk 5 percent of the time.

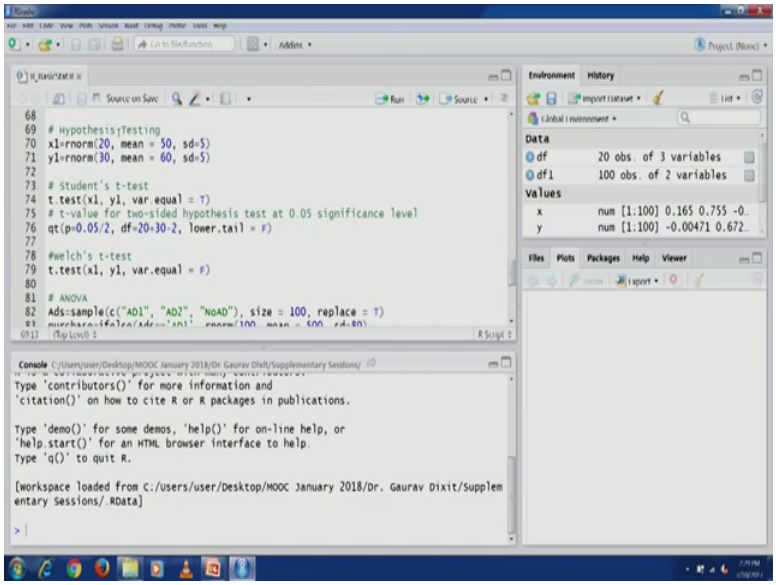
(Refer Slide Time: 07:44)



Now, another term that you might have come across is called p-value; p-value is sum of probability of t being less than or equal to minus absolute of observed t value and probability of t observed t value being greater than or equal to magnitude of observed t value. So, summation of these 2 terms; summation of these 2 numbers is actually going

to be p-value. Now, let us open R studio and let us go through 1 example, which is related to student's t test.

(Refer Slide Time: 08:21)



So, let us first create this hypothetical data, so we have these 2 variables x1 and y1. So, we are going to use R norm command that we discussed in the previous lecture. So, R norm again we want 20 observation with mean, mean 50 and standard deviation being 5 and y1 this is corresponding to the second population we have, we want 30 observations here and the mean being 60 and a standard deviation value being again 5.

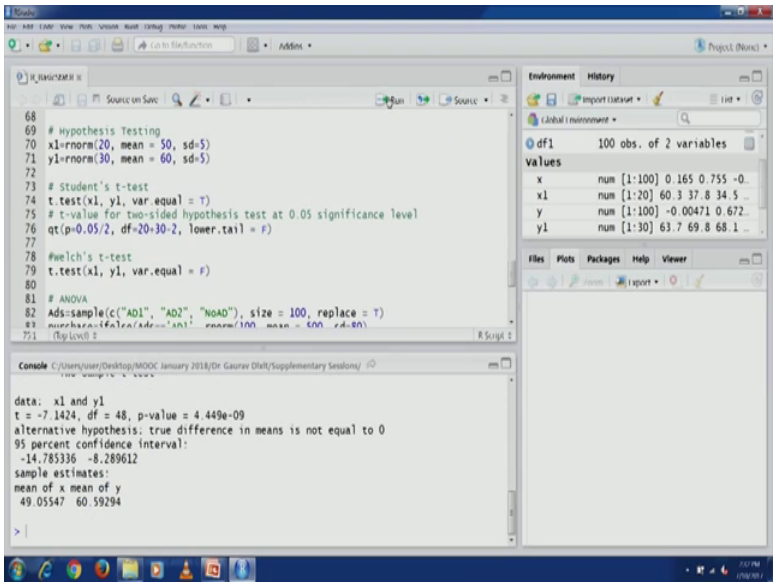
So, you can see that because of the assumptions that are related to student's t-test, you can see we have kept the standard deviation value as same while creating these 2 populations or samples. So, let us compute this, so x1 you would see that x1 has being created here 20 observations and these values are randomly generated and following normal distribution. Now, second sample we can let us execute y1 and will get the 2nd sample you can see in the data section, in the environment section you have y1 30 observation in this here and again the values being generated and again following a normal distribution.

Now, let us come to students t-test. Now, we have this t dot test of function that is available in R and it could be used to run our students t-test. Now, in this case the t test function we pass on x1, that is the first sample and the y1 that is being the second sample and you would see there is another argument called variance var dot equals, this is

related to variance, where as we understand that in this students t-test variance of 2 population are supposed to be equal. So, in this case var dot equal is assigned as true.

So, once you write this particular code we can execute this, so let us do t-test here.

(Refer Slide Time: 10:27)



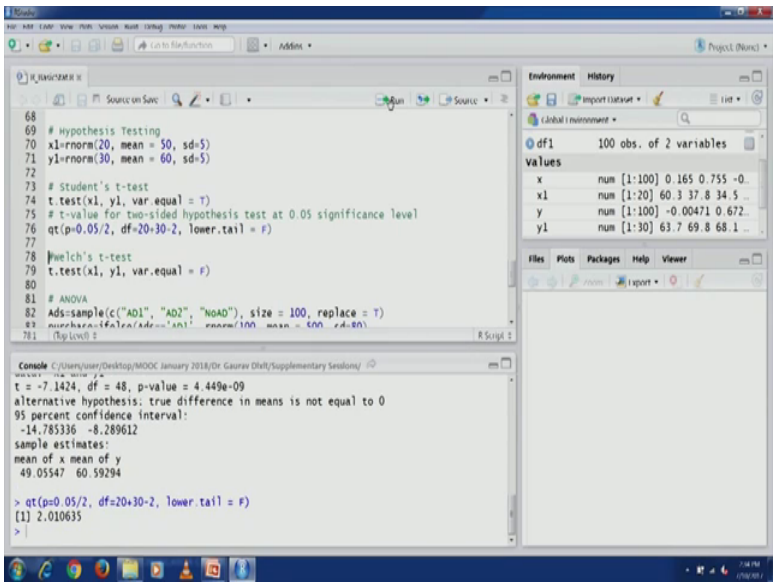
You would see in the result, it is 2 sample t-test because we are trying to compare the means of 2 samples x1 and y1. So, it is 2 sample t-test and you can see the data mention as x1 and y1, you will also get a t statistic t as minus 7.1424, degree of freedom is also mentioned as 48 you can see, that number of observation in x1 sample were 20 and number of observation in y1 sample were 30. So, therefore, addition being 50 and then we subtracting 2 for parameters for mean. So, therefore, it comes out to be 48.

Now, you can also see a p-value has also being computed and how this computation is actually done that we have discussed in the slide. Now, you would also see that alternate hypothesis 2 difference in the result is given there the 2 difference in means is not equal to 0. So, alternate hypothesis is true and the null hypothesis has being rejected in this particular case, you would also see that 95 percent confidence interval has also being mentioned there being between minus 14 to minus 8.289, will have a discussion on confidence interval as well, later in this particular lecture.

Now, you would see mean of x and mean of y those values also being given there. Now, if you want to compare our results of students t-test with t-value which is related which

is t-value corresponding to 0.05 significance level, especially for 2 sided hypothesis test or sample hypothesis test, so we can do this using this particular qt function. So, qt function can give us this value for example, in this case the significance level is 0.05. So, this is going to be divided by 2, being because the normal distribution being symmetric. So, we need to divide because there are going to be 2 regions and this particular value has to be divided for equally for each region and you would also the degree of freedom as 48; 20 plus 30 minus 2 and the lower tail is false. So, once you we can find out the t-value for 0.05, significance level.

(Refer Slide Time: 13:07)



So, let us execute this code and you would find that t value for this given significance level of 0.05 and given degrees of freedom 48 in this case, t-value comes out to be 2.01. Now, if you compare this with the observed t-value that we just saw that is minus 7.1424 it is quite less. So, therefore, the null hypothesis actually rejected.

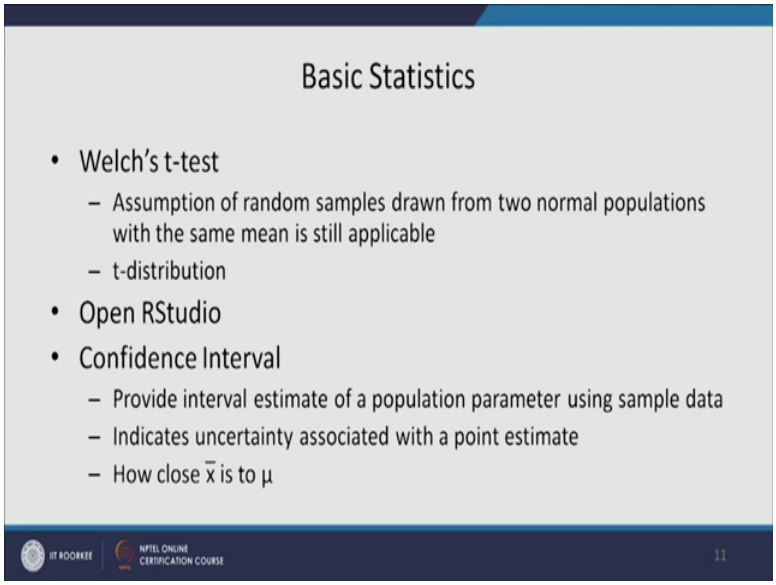
Let us go back to our discussion. Now, another test that can be performed while you know while hypothesis testing related to difference of means is Welch's t-test. So, when do we use Welch's t-test when the population variance are not equal. So, assumption of equal population variance and that is not reasonable and that cannot met and then probably we can use Welch's t-test and do our hypothesis testing. So, again formula for t-statistics for Welch's t-test is given here, so t w is again a difference between 2 samples

x1 bar and x2 bar and divided by, in this case you would see that sample variance $S1^2$ /n1 + $S2^2$ square sample variance for 2nd sample $S2^2$/ n2 and square root has been taken.

So, this is the formula for computing t-statistics. Now, as far as the interpretation is concerned as we did, as we discussed for students t-test again here also the sample means are in the numerator, we have sample means difference of sample means. So, again the same points are applicable, if the value numerator is 0 then probably the null hypothesis is true and if the there is numerator is far from 0 and for low probability like 0.05 is there then probably null hypothesis is good actually be rejected. So, from that sense interpretation of results are going to be same. So, only 1 important difference being that population variance that cannot be assumed as equal.

Now, another assumption that was applicable in student's t-test that random samples would be drawn from normal distributed population that is still applicable. So, again this t-statistics; Welch's t-statistics also follows t-distribution which is as we discussed is very similar to normal distribution and becomes almost normal distribution and the degrees of freedom reach 30 or more.

(Refer Slide Time: 15:39)



So, let us do a small example for Welch's t-test, so let us open R studio. So, again we can use the same data in this for to perform test related to Welch's t-test as well. You can see the t, t dot as the same function can again be used to perform this particular test. Now, the only difference being that variance dot equal now in this case would be assigned as

false. So, you can see 3rd argument var dot equal is false and the 2 samples x1 and y1, we are passing on the same samples and doing the test on the same samples.

So, let us execute this particular line, again you can see here that now the name has changed to Welch's 2 sample t-test; 2 sample being x1 and y1. So, that you can see, you can also see the t-statistics that is Welch's t-statistics that comes out to be minus 6.6412 and degree of freedom comes out to be 30.926, degree of freedom computation in Welch's t-test is slightly different from students t-test.

So, we would not go into detail of that, p-value again the interpretation and meaning remains same. So, here also you will get a p-value. Now, which again this p-value is also less than that low probability value that we talked about less than 0.05, alternate hypothesis 2 difference in means is not equal to 0, again in this case also the null hypothesis is rejected, you can also see in the results 95 percent confidence interval, values are mentioned there, so minus 15 and minus 7.99, more discussion on confidence interval will do in a while in this lecture.

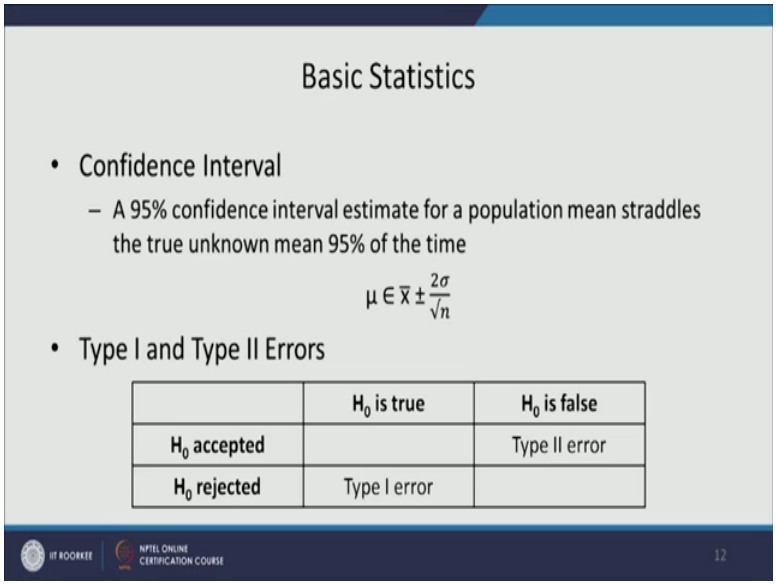
Now, means of these 2 samples; sample estimates is also sample estimates are also given, mean of x and mean of y are also given. Now, if we go back to the earlier computation that we performed about that t-value that is corresponding to 0.05 significance level and the degrees of freedom that computation we can again do and will find out that observed t-value is less than the corresponding t-value where 0.05 significance level and given degrees of freedom in this case, that computation can be done and we will find out that the this particular t is statistics minus 6.6412 is less than that, therefore, null hypothesis has to be rejected.

So, let us go back and so next discussion point is on confidence interval. So, confidence interval actually provides an interval estimate of a population parameter using sample data. So, till now what we are looking for actually the point estimate, but using confidence interval we can also provide an interval estimate of a population parameter. Now, this confidence interval in a way also tells about the uncertainty that is associated with the point estimate. So, point estimate might not be accurate and the confidence interval in a way is explaining this the uncertainty.

Now, another way of understanding confidence interval is how close x bar is to mu. So, that is another way of understanding confidence interval because our x bar is actually

computed based on the sample randomly drawn from the normal distributed population. Now, confidence interval will give us some sense of, will minimize some uncertainty about the sample estimate that we have and it will tell us and range where this particular where we can with some confidence we would be able to say that population parameter is going to lie in that particular range.

(Refer Slide Time: 20:11)



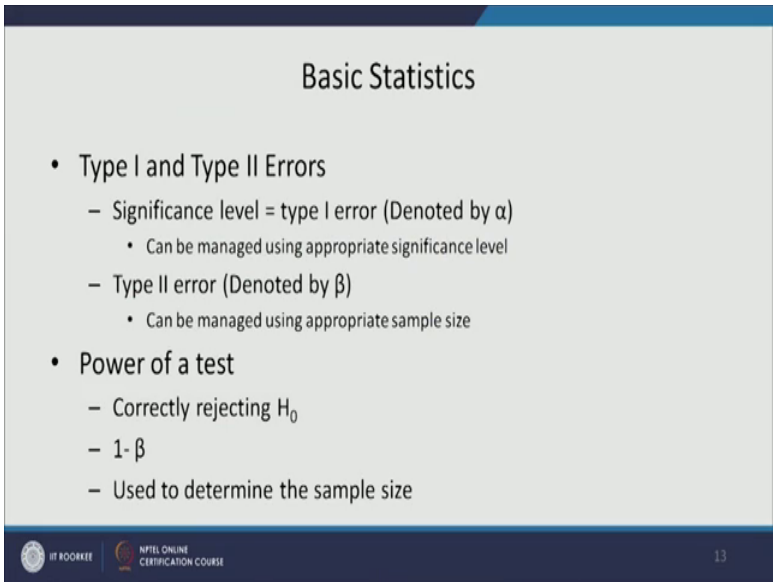
Now, for example, in 95 percent confidence interval estimate for a population mean straddles the true unknown mean 95 percent of the time. Therefore, what we actually mean is that if we are computing an interval estimate based on 95 percent confidence level, then the population mean is 95 percent of the time, the population mean is going to lie in that range. The same thing can be expressed in this form that mu is going to belong to this particular range x bar plus minus twice of sigma divided by square root of n.

So, this particular range, this, using this particular formula range can be computed and the particular population mean will actually straddled by this particular range 95 percent of the time or any other confidence interval, if it is 99 percent confidence interval we are talking about then 99 percent of the time it will straddle that range, similarly for 90 percent. Now, at this point we can discuss 2 important resource related to errors type 1 and type 2 errors. So, is this particular classification table that is displayed in this particular slide, you can see when type 1 error and type 2 error can actually occur. So, for example, if null hypothesis we look at type 1 error if null hypothesis is rejected, while

null hypothesis is being true, so that is called as type 1 error, while the null hypothesis actually true, but it has being rejected using our statistical test or hypothesis test.

The type 2 error occurs when the null hypothesis is false, but using our hypothesis test or statistical test we actually accept null hypothesis. So, these are the 2 situation when type 1 error and type 2 error actually happen, the other 2 are the correct outcome when our hypothesis test accept null hypothesis and is also true or our hypothesis test rejects null hypothesis and is also is false. Now, how do we overcome the problems related to type 1 and type 2 errors. So, for type 1 error we can look at the significance level that is denoted by alpha.

(Refer Slide Time: 22:44)



So, we can manage this particular error using appropriate significance level. So, we reduce the alpha, then there is less chance of doing type 1 error. So, therefore, many times you would see many researcher would prefer 99 percent confidence interval over 95 percent that because they do not want to commit type 1 error. So, therefore, they reduce the alpha depending on their acceptance level. In some research stream or research domain even 90 percent confidence interval is accepted, but in that case there is a risk of committing a type 1 error.

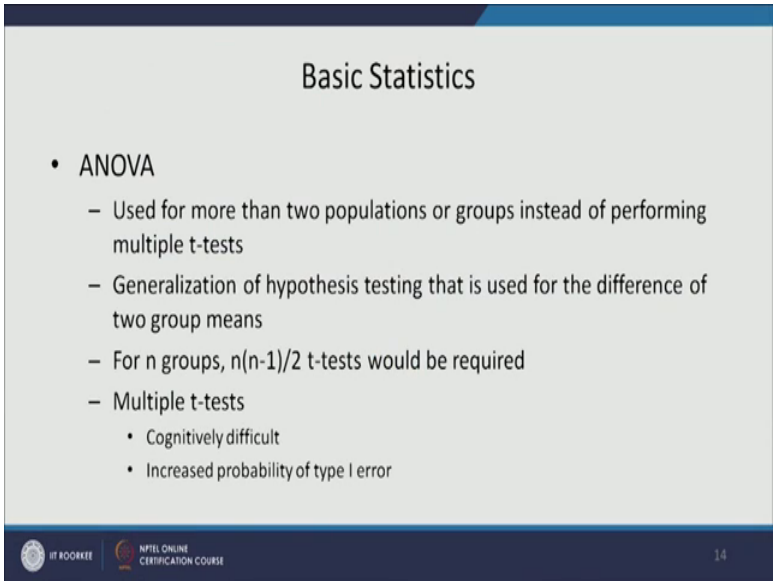
Now, type 2 error; it is generally denoted using beta, this can be generally managed using appropriate sample size. So, if you keep on increasing your sample size and it there is some sort of saturation that is reached and then in that case less chance of committing

type 2 error. Now, another point related to hypothesis testing is power of a test. So, what is a power of a test? So, power of a test is about correctly rejecting null hypothesis.

So, if you go back to the table that we had. So, if you look at null hypothesis is rejected the second row in that case the only problem when the null hypothesis should actually we rejected and it is not is actually the because of the type 2 error. So, you would see that power of a test is actually computed using 1 minus beta because whenever there is type 2 error, then that reduces the power of a test in terms of correctly rejecting null hypothesis. Therefore 1 minus beta is called the power of a test.

So, this particular power of a test is also used to determine the sample size because as we talked about 1 way to manage or handle type 2 error is the selecting appropriate sample size. So, we want to again compute or find out what would be the appropriate sample size, then power of a test could be a good indicator.

(Refer Slide Time: 25:23)



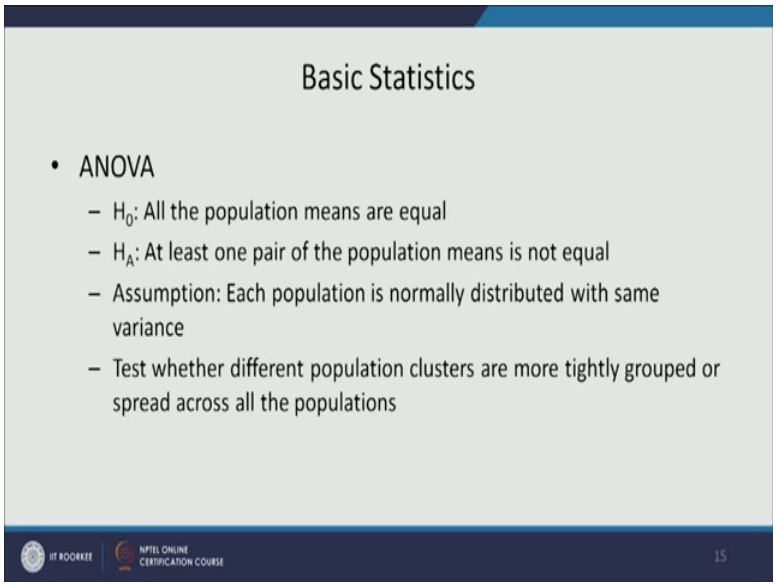
Now, next important statistical test is ANOVA. So, till now what we have been talking about is was mainly about 2 populations. So, what happens about hypothesis testing if you are dealing with more than 2 population, so in that case ANOVA is used. So, used for more than 2 populations or groups instead of performing multiple t-test. So, if we have more than two population, 1 alternative is we perform multiple t-test pairwise t-test for different grouping. So, that is 1 solution, but this can be cumbersome and the interpretation could be cognitively very difficult for us and the probability of committing

type 1 error would actually also increase because when you are trying to do multiple t-test. So, for every pair you have to do interpretation and then it will be influenced by some other t-test and it will become very difficult for you to cognitively interpret the results and the, reduce the manage the probability of committing type 1 error.

So, therefore, ANOVA is preferred in case more than 2 populations are involved. So, another important point related to ANOVA is this is sort of generalization of hypothesis testing and that used for the difference of 2 group means. So, hypothesis testing that is used for difference of 2 group's means it is ANOVA is in a way generalization of that process. So, if we were to perform multiple t-test for n group, then we have to actually do n times n minus 1 divided by 2 test to actually make any conclusion.

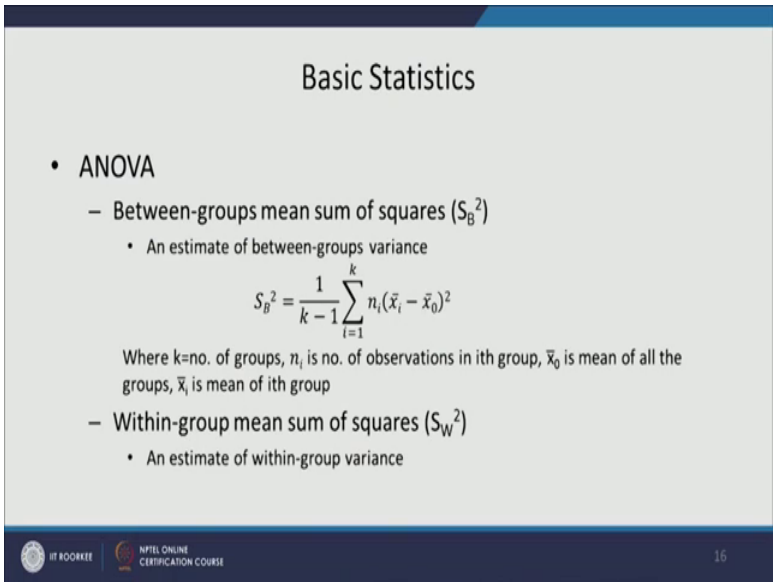
Now, the typical null hypothesis and alternative hypothesis in ANOVAs case is that in ANOVA we assume we assume that in null hypothesis all the population means are equal, which is quite similar to what we do in difference of means, alternate hypothesis at least 1 pair of the population means is not equal.

(Refer Slide Time: 27:35)



So, again assumption is quite similar each population is normally distributed with same variance and the testing is mainly about whether different population; different population clusters whether they are more tightly grouped or spread across the populations, so this is what we are trying to find out.

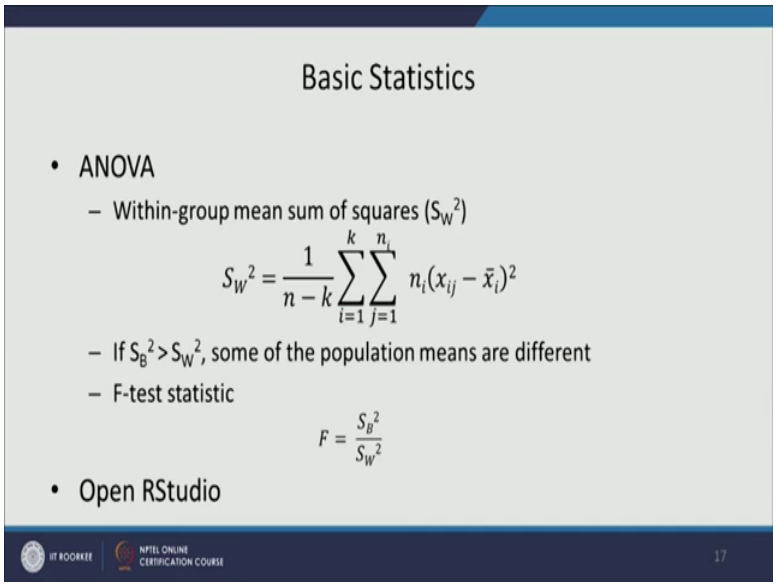
(Refer Slide Time: 28:18)



Now, there are 2 important statistics that we compute in ANOVA process, one is between groups, mean sum of squares that is SB squares. So, this is an estimate of between groups variance, this is the formula SB square can be computed 1 divided by k minus 1, when k being the number of groups and then summation over 1 to k and multiplied by n i, n i is the number of observation in ith group and then the difference between mean of ith group and the mean of all the groups and square of this particular value. So, this is the formula for SB.

So, this is mainly for between group variance, then another estimate that is required related to within group mean sum of a square, it is called within group mean sum of a square and this is an estimate of within group variance. So, we are trying to find out the homogeneity within a group and it is heterogeneity with respect to other groups. So, within group variance is computed in this fashion SW it is called sw, SW^2 is computed in this fashion 1/n-k and then summation over different add k groups 1 to k and the for all the observation from 1 to n i, this is for ith group. So, ni (x ij - xi bar)^2.

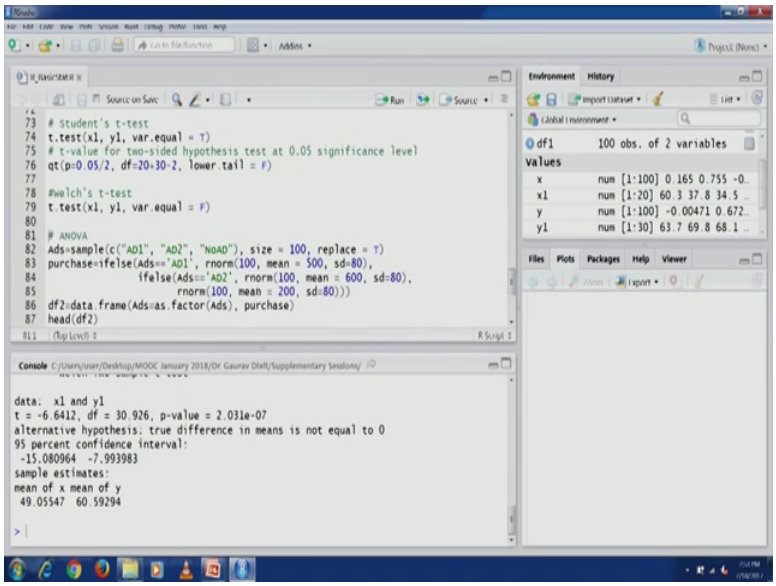
(Refer Slide Time: 29:32)



So, once these 2 computations are done between group mean sum of squares and within group mean sum of square has being computed, then we compare these 2 statistics if SB square is greater than SW square, then we can say that some of the population means are different. So, therefore, null hypothesis would actually be rejected in this case.

So, this is actually done using F-test statistics. So, generally we can use this particular formula F SB square divided by SW square and then the F-test statistics is actually used to find out whether the null hypothesis is accepted or rejected. So, let us go through a small example for ANOVA. So, let us open R studio. So, again in this case we have created a hypothetical data.

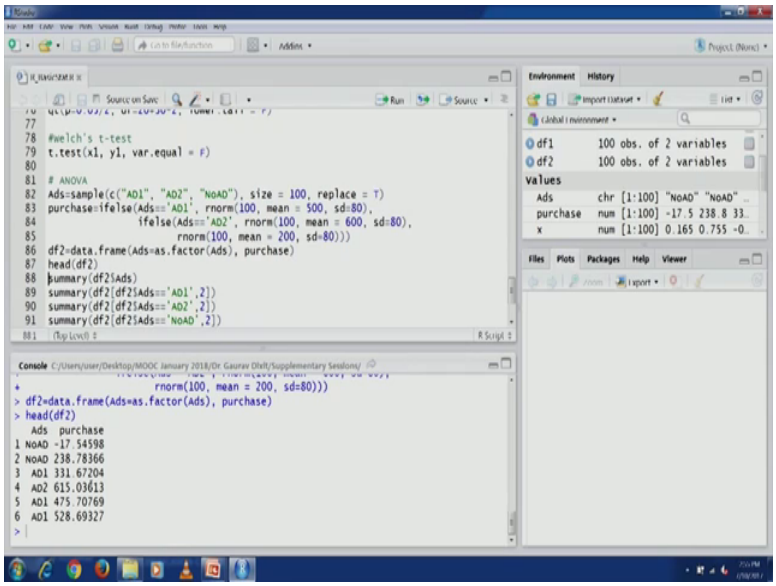
(Refer Slide Time: 30:39)



So, we are talking about ads. So, these are 3 options AD1, AD2 or NOAD at all and the purchase that can be associated with these ads. So, we are again for these 3 types of situation AD1 and AD2 and NOAD, we are trying to create 3 samples randomly again we are using rnorm function, you can see here. So, let us execute first create this particular variable ads, you will see a character vector of ads has being created. So, the sample size is 100.

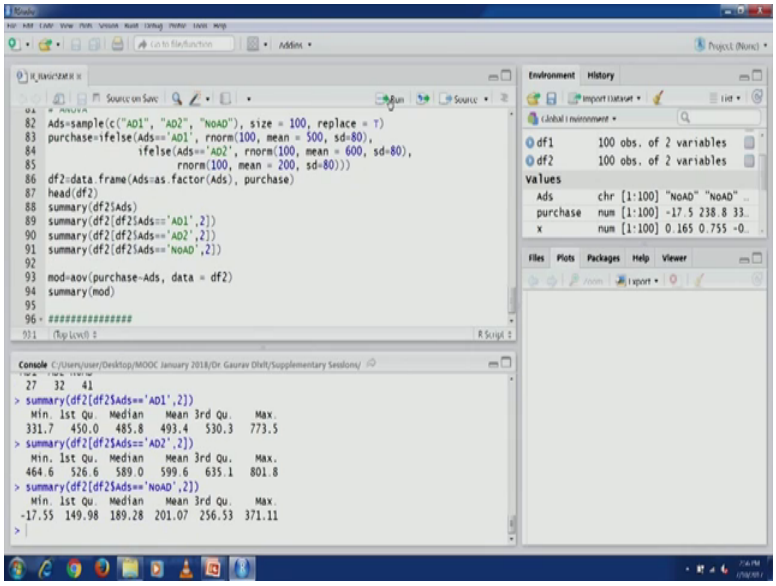
So, this you can see, then purchase we can compute we want to compute you know, we want to generate 3 samples you can see 100 first sample, which is corresponding to ad 1 it is about 100 observation and mean is there 500 standard deviation is there, then for AD2 again 100 observation mean is 600 and standard deviation being 80, you can see that standard deviation is same because that is part of the assumption that variance should be equal, then the 3rd is NOAD case, there also we have 100 observation mean being 200 and standard deviation again is same as previous 2 samples. So, let us execute this particular code, so you would see purchase has being created. Now, we can create a data frame of these 2 variables ads and purchase, so let us do this. So, this is how our first 6 observation of data looks like, so Ads. So, NOAD, NOAD, so these are some of the records and the corresponding purchase value it is also given.

(Refer Slide Time: 32:14)



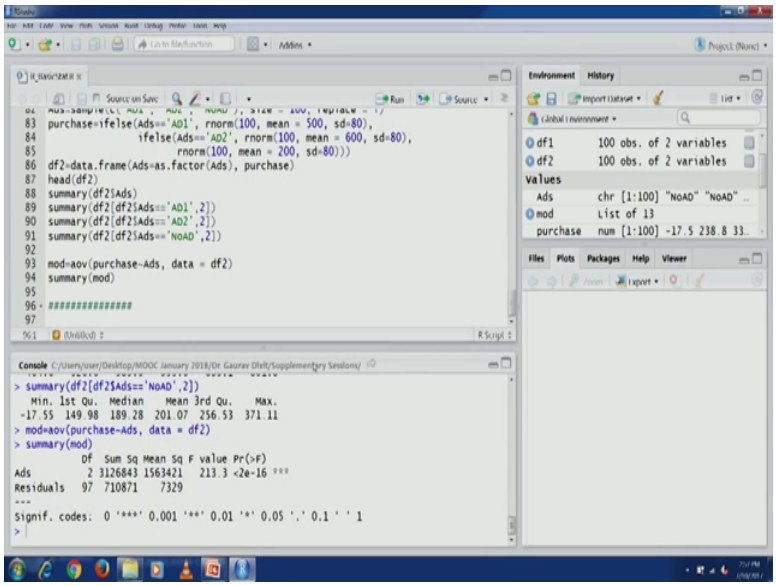
Now, if you are interested in summary you can see AD1, AD2 and NOAD, how they have they are distributed 27, 32 and 41, these are the split for 100 observation. Similarly, if you are interested in statistics related to AD2, specially the purchase part you can see that mean purchase is 493 and the min and max value are also there, similarly we can do the same exercise for AD2, we can find out the statistics related to purchase with respect to AD2, similarly for NOAD situation we can see.

(Refer Slide Time: 33:15)



Now, we have this aov, ANOVA aov to perform ANOVA test. So, in this case you can see first argument is actually a formula. So, in this case purchase is being tested with this respect to ads and data is again df 2, we have data frame 2, we have just created. So, let us run this particular thing; now, let us look at the results.

(Refer Slide Time: 33:46)



You can see in this case F value is there, probability value is there, in this case you would see that because F value is greater than 1, you can see that null hypothesis is rejected, you would also see other numbers here sum of a square and mean squares, so those numbers are here. So, this is how, we can actually perform ANOVA test. So, with ANOVA, we are able to cover the basic statistics using R for this particular course.

Thank you.