

Essentials of Data Science with R Software-1
Professor Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology Kanpur
Lecture 68

Two Sample Test for Mean with Known and Unknown Variances

Hello friends. Welcome to the course Essentials of Data Science with R Software-1 in which we are trying to understand the basic concepts of probability theory and statistical inference. So, you can recall that in the last two lectures, we had considered the test of hypotheses for the mean parameter or the μ from a normal population when σ^2 is known, and when σ^2 is not known. And we have discussed that how are we going to use them how are we going to interpret the results from the software and so, on.

And in software also we had considered that the software command was very general and I had told you that that command we are going to use at that time only for one sample and say, there are some more components which are needed for our two sample test. So, in this lecture, we are going to talk about the two sample test for the mean, when the variance is known and variance is unknown. So, now in this lecture, it is very important that the last two lectures for the cases $H_0: \mu$ equal to μ_0 when σ^2 is known and unknown, you must be wise them because I am simply going to use the same command, same everything.

The only thing what I want to tell you here that under what type of condition these tests can be used and what is the test statistic. So, in this lecture, we are now going to talk about two samples. Whenever in real life we are trying to do the data analysis, there are many question in which we want to compare, for example, on a shopping website, we would like to compare the shopping behavior of say male versus female or we can say that whether these youngsters are doing more shopping or elder people are doing more shopping or we can also say that during a festival time does sales related to the clothing and the sales related to the gift items.

They are going to be the same or they are going to be different on I have different types of questions that are I believe that during the festival time, the sale for the clothing is going to be higher. So, I suggest my shopping website you try to give some coupons or some incentive for the customers or you give better discount because it is possible that not one but a couple of shopping websites are going to float the sale.

So, I working for my website so I would be interested in that thing. So, suppose if there are three websites and they are trying to give the mega sale, bumpers sale etc. etc. during the same period. Now, the customer will have an option whether the customer wants to buy a dress, whether that dress has to purchase from a shopping site 1, 2 or 3. So, obviously, the one shopping site which is giving a better price, better discount for the same item the customer will prefer from there. So how to make comparisons of such things? So, that is that topic, which I am going to discuss in this lecture. So, let us begin our lecture and first try to understand the basics.

(Refer Slide Time: 03:46)

Testing Hypothesis for Difference of Means (Independent Samples: Known Population Variances)
We want to conduct statistical Hypothesis testing of difference of means

- In a survey of buying habits, 400 women shoppers are chosen at random in a super market 'A'. Their average weekly food expenditure is Rs. 250 with a s.d. of Rs. 40.
- For 400 women shoppers chosen at random in some other super market 'B', the average weekly food expenditure is Rs. 220 with a s.d. of Rs. 55.

2

So, now, the first case I will take that we have a situation where σ^2 is known and after that, I will come to the assumption that when σ^2 is unknown. So, you can recall that in the one sample case, you have used the test that was $\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ and then you have conducted the test of hypothesis. So same thing I am going to use here towards the end.

So, now, we are more interested in conducting our test of hypothesis about the difference of the means or comparing the means, for example, in our survey of buying habits 400 women shoppers are chosen at random in a supermarket A and their average weekly food expenditure is suppose Rupees 250 with a standard deviation of Rupees 40 and then there is another supermarket where means another set of 400 women shoppers are chosen at random and their

average weekly food expenditure is coming out to Rupees 220 and the standard deviation here is Rupees 55.

(Refer Slide Time: 04:48)

Testing Hypothesis for Difference of Means (Independent Samples: Known Population Variances)

- Do these two populations have similar shopping habits.
- Are the average weekly food expenditure of two populations of shoppers equal.

Testing Hypothesis for Difference of Means (Independent Samples: Known Population Variances)

We want to conduct statistical Hypothesis testing of difference of means

- In a survey of buying habits, 400 women shoppers are chosen at random in a super market 'A'. Their average weekly food expenditure is Rs. 250 with a s.d. of Rs. 40.
- For 400 women shoppers chosen at random in some other super market 'B', the average weekly food expenditure is Rs. 220 with a s.d. of Rs. 55.

So, how you can conclude that what is really happening in this phenomenon? Do these two populations have similar shopping habits means both these shopping stores A and B, they have got the similar shopping behavior for the woman shoppers or for the female shoppers? Or are the average weekly food expenditure of the two populations are equal because they are trying to

spend her Rupees 250 and Rupees 220 rupees but their variances are also different. You can see here 40 and 55²s.

(Refer Slide Time: 05:18)

Testing Hypothesis for Difference of Means (Independent Samples: Known Population Variances)

- Suppose we have two samples *independent*
- First sample is drawn from a population with mean μ_1 and variance σ_1^2 .
- \bar{x}_1 is the mean and n_1 is the size of 1st sample.
- Second sample is drawn from a population with mean μ_2 and variance σ_2^2 .
- \bar{x}_2 is the mean and n_2 is the size of 2nd sample.
- We wish to examine if two population means μ_1 and μ_2 are different.

So, now that is the question. So, now we have a setup here where we have suppose two samples and they are actually independent sample. Independent means both the samples have been drawn independently from two different population. So, the first sample is drawn from a population with same a mean μ_1 and a variance σ_1^2 and the second sample is drawn from a population with mean μ_2 and variance σ_2^2 .

So, what we try to do here, that from the first sample we are drawing a sample of size n_1 and from the second sample we are trying to draw a sample of size n_2 . So, based on this n_1 and n_2 observations, we try to compute their sample mean in their respective sample. So, the sample mean in the first sample is indicated by \bar{x}_1 and in the second sample the sample mean is indicated by \bar{x}_2 and we wish to examine if the two population mean μ_1 and μ_2 are the same or they are different.

(Refer Slide Time: 06:18)

Testing Hypothesis for Difference of Means (Independent Samples: Known Population Variances)

- To test,

- $H_0: \mu_1 = \mu_2$	$H_1: \mu_1 \neq \mu_2$
- $H_0: \mu_1 = \mu_2$	$H_1: \mu_1 > \mu_2$
- $H_0: \mu_1 = \mu_2$	$H_1: \mu_1 < \mu_2$
- Since population means μ_1 and μ_2 are unknown, we use the statistic $(\bar{x}_1 - \bar{x}_2)$ to make some conclusion about $(\mu_1 - \mu_2)$.
- Assumptions:
 - σ_1 and σ_2 both are known
 - Both populations are normal.
 - Or, sample sizes n_1 and n_2 are large.

$H_0: \mu_1 = \mu_2$
 $H_0: \mu_1 - \mu_2 = 0$
 In general
 $H_0: \mu_1 - \mu_2 = d$

$E(\bar{x}_1 - \bar{x}_2) = E(\bar{x}_1) - E(\bar{x}_2)$
 $= \mu_1 - \mu_2$
 under normal dist.

So, now, I can translate this requirement in terms of the null hypothesis and alternative hypothesis. So, there can be three possible alternative hypothesis against the null hypothesis $H_0 : \mu_1 = \mu_2$. So, this alternative can be you know that two sided or one sided that a $\mu_1 \neq \mu_2$, $\mu_1 > \mu_2$ or $\mu_1 < \mu_2$.

Now, these things are very simple for you to understand. And since this population means μ_1 and μ_2 are unknown so, we use the statistics $\bar{x}_1 - \bar{x}_2$ to make conclusion about $\mu_1 - \mu_2$. You can see here that if you assume the normal population then expected value of $\bar{x}_1 - \bar{x}_2$ is going to be say your $E(\bar{x}_1) - E(\bar{x}_2)$, this is going to be here $\mu_1 - \mu_2$ under normal distribution.

So, you can see here that $\bar{x}_1 - \bar{x}_2$ can be used as an estimator or rather an unbiased estimator of $\mu_1 - \mu_2$. So, we assume here that σ_1 and σ_2 both are known to us, both the populations are normal or if not normal means, we can also assume that the sample sizes are quite large, the n_1 and n_2 's are quite large. So, one point which you have to observe here that here I am trying to write down here H_0 as $\mu_1 = \mu_2$.

So, this can also be written as $H_0 \mu_1 - \mu_2 = 0$ or if you want to make it here more general, then this can also be written as $\mu_1 - \mu_2$ is equal to some value here d . So, this hypothesis can also be used if the difference between the two population mean is 0 or even any other value. So, this is

also helpful in many-many applications that sometimes you want to see whether the difference between the male and female shoppers in our shopping website is has a difference of at least Rupees 1000 or not or you want to see that you have some hypothesis that you believe that females are spending say Rupees 5000 more on shopping the male shopper and so, on. So, these types of things can be done while here. But now, we try to construct here the test of hypothesis for this thing right.

(Refer Slide Time: 08:43)

Testing Hypothesis for Difference of Means (Independent Samples: Known Population Variances)

- We know $\bar{x}_1 \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right)$ and $\bar{x}_2 \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$ → *independently drawn samples*
- Test Statistic:
 $E(\bar{x}_1 - \bar{x}_2) = \mu_1 - \mu_2$
 $E(\bar{x}_1 - \bar{x}_2) = (\mu_1 - \mu_2)$ and $\text{Var}(\bar{x}_1 - \bar{x}_2) = \left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$
Var(\bar{x}_1) + Var(\bar{x}_2) - 2Cov(\bar{x}_1, \bar{x}_2) = 0
- Thus $Z_c = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ when H_0 is true.
 $\mu_1 - \mu_2 = 0$ $H_0: \mu_1 - \mu_2 = d$
- Under our assumptions, $Z_c \sim N(0, 1)$.
- We use $N(0, 1)$ distribution to get critical values and p -values.

So, now, we know that we have drawn a sample from normal population. So, I can write down here that the distribution of the sample mean 1 it is going to be here normal with mean μ_1 and variance σ_1^2/n_1 and sample mean 2 has got a normal population with mean μ_2 and variance σ_2^2/n_2 . Now, based on that we can create the statistics.

So, we know that both is \bar{x}_1 and \bar{x}_2 they are independently drawn, both the samples are independently drawn. So, what will happen here that if you try to see that the observations of first sample and observations of second sample, they are going to be mutually independent of each other.

So, now, in case if I want to estimate here $\mu_1 - \mu_2$, then possibly I can consider $\bar{x}_1 - \bar{x}_2$ as impossible estimator and if you try to take its expectation, you can see here that $\bar{x}_1 - \bar{x}_2$ is an

unbiased estimator of $\mu_1 - \mu_2$ and if you try to find out the variance of this $\bar{x}_1 - \bar{x}_2$, so, this is going to be $\text{var}(\bar{x}_1) + \text{var}(\bar{x}_2) - 2 \text{cov}(\bar{x}_1, \bar{x}_2)$, but this covariance term is going to be 0 because the observations in the first and second sample, they are mutually independent of each other.

So, now this variance of $\bar{x}_1 - \bar{x}_2$ comes out to be like this and you can now write down the statistics that it is just based on the $\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$. So, you can see here I am writing here that I should not that means when x_0 is true that means, when $\mu_1 - \mu_2 = 0$ then I can write down here this statistics $\bar{x}_1 - \bar{x}_2$ and this 0 is coming because $\mu_1 - \mu_2 = 0$.

So, means, if you want to test a hypothesis that H_0 like here $H_0 \mu_1 - \mu_2 = d$, then this 0 is going to be replaced by here d. So, this is just for the information, but here we are considering d to be 0. So, this will become here $\bar{x}_1 - \bar{x}_2 = 0$ divided by the standard deviation which is the positive² root of the variance which you have obtained here. So, now, this is the statistics under H_0 , that means when H_0 is true, so, and we know from our statistical theory that this quantity is going to follow a normal distribution with mean 0 and variance 1 right. So, we use the normal 0,1 distribution to get the critical values and p-values.

(Refer Slide Time: 11:24)

```

Testing Hypothesis for Difference of Means (Independent Samples:
Known Population Variances) in R
Classical Gauss Test

Gauss.test is available in library "compositions"
(Compositional Data Analysis). So it needs to be installed first.

install.packages("compositions")
library(compositions)

Usage
Gauss.test(x, y, mean=0, sd=1, alternative =
c("two.sided", "less", "greater"))

```

So, now, if you want to use this test in our software, that is pretty simple and straight forward, you already have done this test, you have done the Gauss test in the package compositions when

we did the test of hypothesis for mean when σ^2 is known. So, in the same command in the same setup, you simply have to now give here the values for the second sample. Earlier we had taken all such values which were related to the second sample as null. So, we are going to use that command here Gauss dot test for that you need to install the package composition and you need to upload it and after that the command here is Gauss dot test, G is going to be here in capital letter, G-a-u-s-s dot t-e-s-t and now, you have to give her the data vector.

So, x is going to be the data first sample and y is going to be the data from the second sample and then this mean actually mean is the value of $\mu_1 - \mu_2$. So, if this is equal to you have to be careful if it is a $d = \mu_1 - \mu_2$ then in that case you can give this mean to be equal to here d. Some say some numerical value, but in this case, because we are considering the value of $\mu_1 - \mu_2$ to be 0, so, we are giving it to be here 0. Now, a standard deviation you have to just specify an alternative you have to specify whether you want two sided, less or greater and then you can specify here the confidence level also.

(Refer Slide Time: 12:55)

Testing Hypothesis for Difference of Means (Independent Samples: Known Population Variances) : Example in R
Following are the gain in weights (in grams) of fishes fed on two diets A and B.

A: 25, 32, 30, 34, 24, 14, 32, 24, 30, 31, 35, 25

B: 44, 34, 22, 10, 47, 31, 40, 30, 32, 35, 18, 21, 35, 29, 22

Suppose the standard deviation is known to be 2

We want to test if the two diets differ significantly as regards their effect on increase in weight.

$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$

So, anyway we try to use this test on a given set of data and we try to see what we can do here. So, now, we consider here an example, where we try to implement this thing in the R software. So, suppose that there are some fishes and they are given two different types of food two different types of diets A and B and how much weight they have gained, this is reported.

So, from the two samples the gain in the weights of the fish is reported here is like this, for A like this and for B be like this and these weights are given in the grams and suppose the standard deviation is known to be here 2. And suppose we want to test if the two diets differ significantly as regard their effect on increase in the weight. So, we want to test your hypothesis like $H_0 \mu_1$ equal to μ_2 versus $H_1 \mu_1$ is not equal to μ_2 .

(Refer Slide Time: 13:47)

```

Testing Hypothesis for Difference of Means (Independent Samples:
Known Population Variances) : Example in R
xa = c(25,32,30,34,24,14,32,24,30,31,35,25)
xb =
c(44,34,22,10,47,31,40,30,32,35,18,21,35,29,22)

Gauss.test(xa, xb, mean=0, sd=2, alternative =
c("two.sided"))

one sample Gauss-test
data: xa
T = -2, mean = 0, sd = 2, p-value = 0.009823
alternative hypothesis: two.sided

```

Handwritten notes on the slide:
 A, B
 $H_0: \mu_1 - \mu_2 = 0$
 $\alpha = 0.05$
 $p < \alpha$
 Reject H_0

So, for that, I simply have to give here the data here like this, I am connoting here say data on xa as A, data on the food B as xb which is here like this. Now I have to just use that R command that Gauss dot test and then is the data on the food A, xb is the data on the food B, mean is equal to 0 because we are considering $\mu_1 - \mu_2$ to be 0 under H_0 and sd is the standard deviation that is given to be here 2, the alternative we are considering here two sided.

So, now you can see here this is here the outcome. It is just exactly on the same lines as we did in the case of when σ^2 is known in case of μ . So, the data here is given here to be here like this, t will come out to be at -2 mean here is 0, sd here is 2 and p-value here is 0.009823 and alternative here is two sided. So, now you can just compare whether p-value is less than α or not.

So, in case if you try to take your α to be here 0.05, then what will happen this p-value is smaller than α and we will say here reject H_0 . That means $\mu_1 - \mu_2$ is not equal to 0. Now, whether it is

increasing the weight or decreasing the weight that we don't know. For that you have to conduct the other two types of hypothesis.

(Refer Slide Time: 15:14)

```
R Console
> xa = c(25,32,30,34,24,14,32,24,30,31,35,25)
> xa
[1] 25 32 30 34 24 14 32 24 30 31 35 25
> xb = c(44,34,22,10,47,31,40,30,32,35,18,21,35,29,22)
> xb
[1] 44 34 22 10 47 31 40 30 32 35 18 21 35 29 22
> Gauss.test(xa, xb, mean=0, sd=2, alternative = c("two.sided"))

one sample Gauss-test

data: xa
T = -2, mean = 0, sd = 2, p-value = 0.009823
alternative hypothesis: two.sided
```

And this is the screenshot of the same outcome right.

(Refer Side Time: 15:17)

Testing Hypothesis for Difference of Means (Independent Samples: Unknown Population Variances) :

- When σ_1 and σ_2 both are known, we use $N(0, 1)$ distribution.
- When σ_1 and σ_2 both are not known, we use t distribution
- Assumptions:
 - Both populations are normal. ✓
 - Unknown σ_1 and σ_2 are equal. $\sigma_1^2 = \sigma_2^2$
- For large sample sizes, we can still use $N(0,1)$ distribution.

Fisher Bahnan's problem.

And I would like to show you on the art console, but let us first consider the second test, where we are going to simply assume the same setup, but we assume that the variances are unknown to

us. So, now, we have done this case that when σ^2 is unknown in the one sample test for the mean, then we have to use that t statistics. So, similar type of thing we are going to use here also.

So, when σ_1 and σ_2 both are known, we use the $N(0,1)$ distribution. When σ_1 and σ_2 both are unknown, then we use the t distribution and assumptions are both the populations are normal and we are assuming here that σ_1^2 and σ_2^2 are unknown, but equal and there is a reason for that, that whenever we are trying to find the test using our statistical theory, in that test, we need to find the distribution of certain statistics and for which we need to make an assumption that $\sigma_1^2 = \sigma_2^2$ and under that condition, the test statistics what is being reported here that is obtained.

Now, the question comes here that in case if this is not valid, then definitely we have another type of test and that is actually called as Fisher Behrens problem. In that problem, they have tried to address that if σ_1^2 and σ_2^2 are unequal as well as unknown, then in that case, one can conduct the test of hypothesis. But definitely there is going to be some loss in the efficiency.

So, well, I will not discuss here the Fisher Behrens problem or how to construct the test when σ_1^2 and σ_2^2 are unknown, as well as unequal but definitely through the software, this is not a big deal, you simply have to choose the correct option and then based on that, you can conduct the test of hypothesis, but you must know from the basic fundamental, what are you going to do and that is important, because many times people are trying to use the software and they try to give these options without understanding and but software does not understand that those options have been given without understanding. So, if they do not understand these options, possibly the statistical conclusions are going to be wrong.

(Refer Slide Time: 17:45)

Testing Hypothesis for Difference of Means (Independent Samples: Unknown Population Variances) :

- To test,
 - $H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$
 - $H_0: \mu_1 = \mu_2$ $H_1: \mu_1 > \mu_2$
 - $H_0: \mu_1 = \mu_2$ $H_1: \mu_1 < \mu_2$
- Let
 - ✓ $x_1, x_2, \dots, x_m \sim N(\mu_1, \sigma_1^2)$ and
 - ✓ $y_1, y_2, \dots, y_n \sim N(\mu_2, \sigma_2^2)$ *independently*
 $E(\bar{x} - \bar{y}) = \mu_1 - \mu_2$
- Population means μ_1 and μ_2 are unknown, we use the statistic $(\bar{x} - \bar{y})$ to make some conclusion about $(\mu_1 - \mu_2)$.
- σ_1 and σ_2 both are unknown.

So, now, we come to our test of hypothesis. Suppose we want to get the null hypothesis, $H_0 \mu_1$ equal to μ_2 and there are three possibilities for the alternative μ_1 is not equal to μ_2 . μ_1 is greater than μ_1 and μ_1 is less than μ_2 . So, now, what we try to do? We try to observe the two samples independently from to normal population. So, the first sample is x_1, x_2, \dots, x_m of size m . And this is drawn from the normal population with mean μ_1 and variance σ_1^2 and the second sample is here y_1, y_2, \dots, y_n of size small n which is drawn from the normal μ_2 mean and variance σ_2^2 and both are drawn independently.

That is obvious. So, this population mean μ_1 and μ_2 both are unknown, and we want to use the statistics $\bar{x} - \bar{y}$ to make some conclusion about $\mu_1 - \mu_2$ because you know that expected value of $\bar{x} - \bar{y} = \mu_1 - \mu_2$ that means $\bar{x} - \bar{y}$ is going to work as an unbiased estimator for μ_1 and μ_2 and we assume that here obviously σ_1 and σ_2 both are unknown, but they are equal. That is what you have to keep in mind.

(Refer Slide Time: 18:55)

Testing Hypothesis for Difference of Means (Independent Samples: Unknown Population Variances)

- We know $\bar{x} \sim N\left(\mu_1, \frac{\sigma_1^2}{m}\right)$ and $\bar{y} \sim N\left(\mu_2, \frac{\sigma_2^2}{n}\right)$
- Test Statistic:
 $E(\bar{x} - \bar{y}) = (\mu_1 - \mu_2)$ and *Pooled estimate of variance*
 $S^2 = \frac{1}{m+n-2} \left[\sum_{i=1}^m (x_i - \bar{x})^2 + \sum_{j=1}^n (y_j - \bar{y})^2 \right]$
 *$\frac{1}{m-1} \sum_{i=1}^m (x_i - \bar{x})^2 + \frac{1}{n-1} \sum_{j=1}^n (y_j - \bar{y})^2$
 $\frac{(m-1)s_x^2 + (n-1)s_y^2}{(m-1) + (n-1)}$*

Thus $T_c = \frac{(\bar{x} - \bar{y}) - 0}{S \sqrt{\frac{1}{m} + \frac{1}{n}}} = \frac{(\bar{x} - \bar{y})}{S \sqrt{\frac{1}{m} + \frac{1}{n}}}$ when H_0 is true.
 $\mu_1 - \mu_2 = 0$

- Under our assumptions, $T_c \sim t_{m+n-2}$ *Choose the values do t from t tables*
- We use $N(0, 1)$ distribution to get critical values and p-values.

So now, from the results that we did in the lecture of distribution of sample means, we know that the sample mean \bar{x} is going to follow our normal distribution with mean μ_1 and variance σ_1^2/n upon n , and sample mean of y_1, y_2, \dots, y_n that is \bar{y} , this is going to follow a normal distribution with mean μ_2 and variance σ_2^2/n .

And based on that, we can create the test statistics as follows, we know that expected value of $\bar{x} - \bar{y}$ is going to be $\mu_1 - \mu_2$ that I already have shown you. But now for estimating the variance what we try to use here, we try to use here the pooled estimate of variance. So what I am going to do here, because I like this one that you try to compute the

$$S^2 = \frac{1}{m+n-2} \left[\sum_{i=1}^m (x_i - \bar{x})^2 + \sum_{j=1}^n (y_j - \bar{y})^2 \right].$$

So, this is actually something like $((m-1)s_x^2 + (n-1)s_y^2)/(m+n-2)$ and then this $m+n-2$ is obtained like as your $(m-1) + (n-1)$. So and this is s_x^2 and s_y^2 they are the quantity of like $1/(n-1) \sum_{i=1}^m (x_i - \bar{x})^2$ and s_y^2 is $(1/(n-1)) \sum_{j=1}^n (y_j - \bar{y})^2$. So, now using this pool estimator of variance, we can define here the statistics T_c . So, that is going to be simply $\bar{x} - \bar{y}$ and this 0 is coming because of this here $H_0 : \mu_1 - \mu_2 = 0$.

If you want to do something else as we discussed earlier, you can also take here as d and then you can conduct the relevant test of hypothesis and then in the denominator this is capital S , which is the positive square root of this s square and the square root of 1 upon n_1 and this is the square root of 1 upon n plus 1 upon n . So, this can be written here as like this $\frac{(\bar{x} - \bar{y})}{S \sqrt{\frac{1}{m} + \frac{1}{n}}}$ and this is

the value of the T_c when H_0 is true because when H_0 is true only when $\mu_1 - \mu_2$ is taking the value 0 , otherwise this will take some other value say d . So, under this assumption, we can find out the distribution of this statistics at t distribution with $m + n - 2$ degrees of freedom.

And in case if you are trying to use a larger sample also, which will usually be a case in the data science, then we are going to obtain the critical values from the $N(0, 1)$ distribution, because usually you are you have a condition only where the sample size is greater than 30 or the degrees of freedom should be greater than 30 . So, that is really going to happen in data science. Otherwise, in cases, if it is not another case, then you please try to choose the values of t from t tables, the probabilities of t tables. That that is what you have to do and this is what I wanted to write here.

(Refer Slide Time: 22:15)

Testing Hypothesis for Difference of Means (Independent Samples: Unknown Population Variances): Example in R
 Following are the gain in weights (in grams) of fishes fed on two diets A and B.

A: 25, 32, 30, 34, 24, 14, 32, 24, 30, 31, 35, 25

B: 44, 34, 22, 10, 47, 31, 40, 30, 32, 35, 18, 21, 35, 29, 22

Suppose the standard deviation is unknown.

(Note that now the sd is unknown (earlier it was known))

We want to test if the two diets differ significantly as regards their effect on increase in weight.

$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$

Now, I try to take here the same example and I try to show you how you can do it in the R console. So, we have the two samples in which we have obtained the data on the weights of the

fish based on two diets A and B and so, this data is given here for the diet A and the diet B is here like this and these values are the gain in the weights of the fishes.

So, they are given in grams. Earlier, we assume that a standard deviation was known to be 2 but now we are assuming that the standard deviation is not known to us. So, now we have to estimate it and then we have to compute the t statistic, but we do not have to do anything. Well, if you want, you can do it manually also, but then we are going to use the R software to do the same job. Now, I have given you a sufficient idea that how these values are calculated and how these decisions are made based on left tail test, right tail test or two sided test. So, we want to test the hypothesis same hypothesis $H_0 \mu_1$ equal to μ_2 versus $H_1 \mu_1$ is not equal to μ_2 .

(Refer Slide Time: 23:14)

```

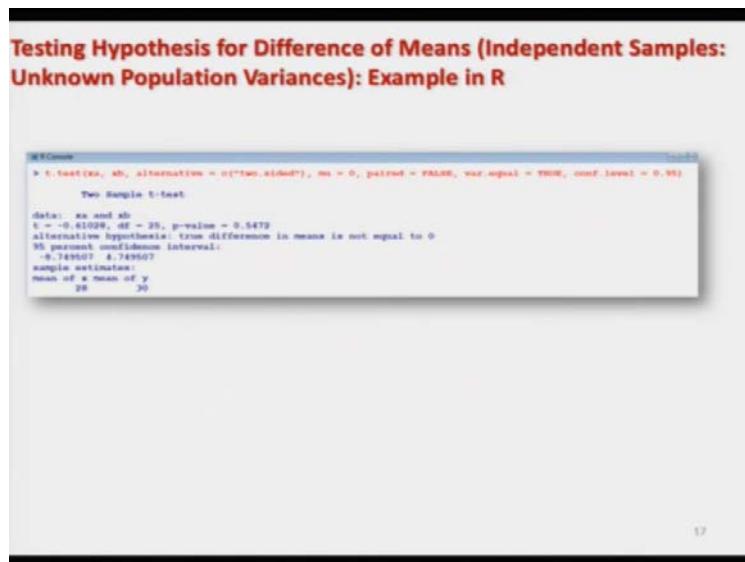
Testing Hypothesis for Difference of Means (Independent Samples:
Unknown Population Variances): Example in R
xa = c(25, 32, 30, 34, 24, 14, 32, 24, 30, 31, 35, 25)
xb = c(44, 34, 22, 10, 47, 31, 40, 30, 32, 35, 18, 21, 35, 29, 22)
t.test(xa, xb, alternative = c("two.sided"), mu = 0,
paired = FALSE, var.equal = TRUE, conf.level = 0.95)
Two Sample t-test
data: xa and xb
t = -0.61028, df = 25, p-value = 0.5472
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
-8.749507  4.749507
sample estimates:
mean of x mean of y
      28      30

```

But now you can see here I have to give here this data like this. I have to simply give here the same data xa and xb as earlier and I have to use the same command but now you have to observe where is going to be the change. The command is the same, so we are going to use the t test, the data on the diet A, the data on diet B. That is xa and xb, alternative here is the same two sided μ is equal to here 0 which is the value of here $\mu_1 - \mu_2$ equal to 0 under H_0 and paired is equal to FALSE because we have not used paired observation, but the difference comes over here that variance dot equal to is now true.

So, this is going to indicate this option that here in this case, the variances are unknown and variances are unequal and then confidence level, you have to give the value of $1 - \alpha$ as earlier and the output is going to be the same, exactly you have to interpret in the same way. This is here the p-value and this is here the 95 percent confidence interval and since this is a two sided hypothesis, so you can see here that these intervals are given here as lower limit and upper limit. Now, I will ask you that please try to look into the earlier values and try to see that whether this confidence interval is going to be changed or not when this standard deviation are known or say unknown.

(Refer Slide Time: 24:37)



```
Testing Hypothesis for Difference of Means (Independent Samples:
Unknown Population Variances): Example in R

> t.test(xa, xb, alternative = c("two.sided"), mu = 0, paired = FALSE, var.equal = FALSE, conf.level = 0.95)

Two Sample t-test

data: xa and xb
t = -0.61028, df = 25, p-value = 0.5472
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.749507  0.749507
sample estimates:
mean of x mean of y
 20          30
```

Now, you can see here, this is the screenshot.

(Refer Slide Time: 24:41)

Testing Hypothesis for Difference of Means (Dependent Samples):

Assume a company send their salespeople to a "customer service" training workshop. They want to know has the training made a difference in the number of complaints.

Following data is collected:

Salesperson	Number of Complaints	
	Before	After
C.B.	6	4
T.F.	20	6
M.H.	3	2
R.K.	0	0
M.O.	4	0

18

Testing Hypothesis for Difference of Means (Independent Samples: Unknown Population Variances): Example in R

```
xa = c(25, 32, 30, 34, 24, 14, 32, 24, 30, 31, 35, 25)
xb = c(44, 34, 22, 10, 47, 31, 40, 30, 32, 35, 18, 21, 35, 29, 22)
```

```
t.test(xa, xb, alternative = c("two.sided"), mu = 0,
paired = FALSE, var.equal = TRUE, conf.level = 0.95)
```

$H_0: \mu_1 - \mu_2 = 0$

```
Two Sample t-test
data: xa and xb
t = -0.61028, df = 25, p-value = 0.5472
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
-8.749507  4.749507
sample estimates:
mean of x mean of y
28         30
```

16

And then I will try to show you this outcome towards the end because now I understand that you are pretty comfortable with the R software. And you can see I simply had to just copy the same command here in the R console to assure you the yes that these things are working. So now this was about the t test when the variances are unknown, but we have imposed a condition that variances are equal.

Now, in case if you have a condition that the variances are not equal, my suggestion is simply that you try to use this option in the command and then simply try to get here variance dot equal

is equal to say here FALSE and that will use the concept from the Fisher Behrens theorem and it will try to give you the outcome, there is no issue. But anyway, I am not going to talk here about the Fisher Behrens problem and the related details.

Now, I try to take here one more example of the t test where the variances are unknown, we have samples from two different populations that is two sample tests, where the variances are unknown. But in this case, the observations are paired. If you remember, we had talked about one sample test two sample tests and in the two sample tests, we are taking an example we are the observations are obtained on the same unit that mean for example, if a student has been given a test and his marks are compared.

When the student has been given a special tuition or training and then the examination is conducted again or the same thing is this that if you want to tell the efficiency or efficacy of a medicine, then some patients are called, their body parameters are recorded, and then they are given the medicine for some time and then the body parameters are recorded again on the same set of patients.

So, these are the situations where we can employ the paired t test. So paired t test is going to be applied when we have two samples, the variances are unknown, but the observations are paired that means on the same unit, you are going to obtain two sets of observation, so the two samples are not independent.

So let me try to take a very simple example and then I try to give you the details. Suppose there is a company in which there are some people who are working in the customer service. So and they have to attend some telephone calls suppose or say these complaints and then they have to attend them. So now, what they do, they try to provide a special training to those people. And then they want to compare whether the number of complaints after the training or before the training is there any change, whether they have become more or less and so on this type of conclusion.

So, this company has a suppose here, say here, these salesperson, there is some name is given C.B, T.F and so on, whatever it is, whatever you want. Now, the main thing for us is that the number of complaints which these people were able to solve before the training, this was like 6,

20, 3 etc. And after the training, the same person attended only 4 calls, the second person attended only 6 calls, third person attended only 2 calls and so on. So, now, they want to test whether that training was effective or not.

(Refer Slide Time: 28:03)

Testing Hypothesis for Difference of Means (Dependent Samples):

- To test,

- $H_0: \mu_1 = \mu_2$	$H_1: \mu_1 \neq \mu_2$
- $H_0: \mu_1 = \mu_2$	$H_1: \mu_1 > \mu_2$
- $H_0: \mu_1 = \mu_2$	$H_1: \mu_1 < \mu_2$
- Let
 - $x_1, x_2, \dots, x_n \sim N(\mu_1, \sigma_1^2)$ and
 - $y_1, y_2, \dots, y_n \sim N(\mu_2, \sigma_2^2)$
- Both samples are dependent samples.

So, that means once again we are interested in the same type of test of hypothesis that is $H_0 : \mu_1 = \mu_2$ and versus H_1 that can be two sided or one sided like greater than or less than that is μ_1 is not equal to μ_2 or μ_1 is greater than μ_2 , or μ_1 is less than μ_2 . So, now, if you try to see in this appeared sample cases situation because we are trying to obtain the observations on the same set of people. So, obviously, the number of observations in both the sample they are always going to be the same. So, we observe here two samples on the x_1, x_2, \dots, x_n and environment y_1, y_2, \dots, y_n from the normal populations and what the samples in this case they are dependent sample. It is unlike the earlier case.

(Refer Slide Time: 28:48)

Testing Hypothesis for Difference of Means (Dependent Samples):

- The two samples are not independent.
- But sample observations are paired together.
- This means the pair of observations x_i and y_i corresponds to the same i^{th} individual.
- Clearly, the sizes of both samples should be the same.
- Population means μ_1 and μ_2 are unknown, we use the statistic $(\bar{x} - \bar{y})$ to make some conclusion about $(\mu_1 - \mu_2)$.
- We obtain the differences $d_i = x_i - y_i$ for each pair of observations.
- Assumption: Both populations are normal.

Now, so, what we have to observe here that two samples are not independent and sample observation are paired together. This means the pair of the observation x_i and y_i , they correspond to the same i^{th} individual. So, obviously, in that in this case, the sizes of both the sample should be the same and the population mean μ_1 and μ_2 are unknown and we use the statistics $\bar{x} - \bar{y}$ once again to calculate about $\mu_1 - \mu_2$ as this is going to be an unbiased estimator.

So, in this case, now, you have to observe what are we going to do. In the first step we obtain the difference between their individual observations and we denote them here, let us say d_i , $x_i - y_i$. Please don't get confused with the same d that we use earlier when I discussed that $H_0 \mu_1 - \mu_2 = d$, this is something else this is different and we assume that here both the propositions are going to be normal.

(Refer Slide Time: 29:45)

Testing Hypothesis for Difference of Means (Dependent Samples):

- Since, we use the differences $d_i = x_i - y_i$ in our analysis. We need the population variance of d_i 's.
- In practice this variance is not available.
- Even if σ_1 and σ_2 are known, then also we can not obtain population variance of d_i 's. (samples are not independent)
- So, this test can be considered as
 - a test of single mean (of d_i 's).
 - with hypothetical mean as zero.
 - population variance unknown.

21

So, now we are not going to use x_i and y_i but we are going to use this d_i to develop the test statistic. So, that is why we need the mean and variance of this d_i and in practice this variance is unknown to us. And even if suppose σ_1^2 and σ_2^2 are known to us, but then finding out the variance of this new random variable $x_i - y_i$ might be complicated, might be different, because the samples are not independent.

Earlier finding odd variants of $x_1 - y_1$ was easier because both the samples were independent, but in this case they are dependent. So, we try to translate this two sample case in the framework of one sample case, how? As follow, we try to consider this case as a test of single mean of d_i 's with hypothetical mean as 0 and population variance as unknown.

(Refer Slide Time: 30:45)

Testing Hypothesis for Difference of Means (Dependent Samples):

Thus the test statistic is

$$T_c = \frac{\bar{d}}{s_d/\sqrt{n}} \text{ when } H_0 \text{ is true.}$$

where $\bar{d} = \frac{1}{n} \sum_{i=1}^n d_i$, $s_d^2 = \frac{1}{n-1} \sum_{i=1}^n (d_i - \bar{d})^2$

- Under our assumptions, $T_c \sim t_{n-1}$.

Handwritten notes:
 $\frac{\bar{d} - \mu}{s_d/\sqrt{n}}$
 $\frac{d_1 - d_n}{\bar{d} - \mu}$
 $\frac{s_d^2}{s_d/\sqrt{n}}$

So, now what you try to see here earlier you had taken the test with like $\frac{\bar{x} - \mu}{s/\sqrt{n}}$. So, what we try to do here, that instead of using here the sample observation x_1, x_2, \dots, x_n , now, we try to use the observation d_1, d_2, \dots, d_n and we try to find out here $\bar{d} - \mu$, which is going to be say here $H_0 : \mu = 0$ under H_0 divided by the standard error of d_1, d_2, \dots, d_n , means the standard error of \bar{d} based on d_1, d_2, \dots, d_n divided by the sample size. This is what exactly are we doing. So, this case is now similar to your one sample t test what you did when the σ^2 is unknown to us.

So, we are simply trying to write down this statistics here T_c this is equal to $\frac{\bar{d}}{s_d/\sqrt{n}}$ when H_0 is true and where this here \bar{d} is going to be as simply the sample mean of d_1, d_2, \dots, d_n and s_d^2 is going to be the sample variance of d_1, d_2, \dots, d_n and under our assumption this statistics T_c is going to follow a t distribution with $n - 1$ degrees of freedom because now, this is compatible with a one sample t test.

(Refer Slide Time: 31:57)

Testing Hypothesis for Difference of Means (Dependent Samples):

Usage

```
t.test(x, y, paired = TRUE)
```

```
t.test(x, y, alternative = c("two.sided",  
"less", "greater"), mu = 0, paired = TRUE,  
var.equal = TRUE, conf.level = 0.95)
```

23

And in case if you want to use the R software to do such a test, then it is very simple. The same command continues there, the only option you have to change here that you have to change the option of paired. Earlier you used paired is equal to FALSE now, you have to use here paired is equal to TRUE. So, in this case, you can see the command remains the same, t dot test everything remains the same, only the option this paired equal to true is going to be changed. So, now, you can see here that here in this case paired is equal to here TRUE.

(Refer Slide Time: 32:33)

Testing Hypothesis for Difference of Means (Dependent Samples):

Arguments

- x** a (non-empty) numeric vector of data values.
- y** a (non-empty) numeric vector of data values.
- alternative** a character string specifying the alternative hypothesis, must be one of "two.sided" (default), "greater" or "less".
- mu** a number indicating the true value of the mean (or difference in means if you are performing a two sample test).
- paired** a logical indicating whether you want a paired t-test.

24

Testing Hypothesis for Difference of Means (Dependent Samples): Example in R

Assume a company send their salespeople to a "customer service" training workshop. They want to know has the training made a difference in the number of complaints.

Following data is collected:

Salesperson	Number of Complaints		Difference, d_i (2-1)
	Before (1)	After (2)	
C.B.	6	4	-2
T.F.	20	6	-14
M.H.	3	2	-1
R.K.	0	0	0
M.O.	4	0	-4

d_1
 d_2
 d_3
 d_4
 d_5

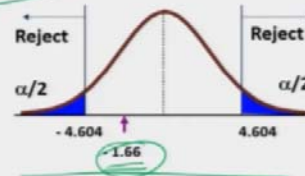
So, now, you can use them on our real data set, but I tried to take here the same example and I try to show you these are the same option that I used that I showed you earlier. So, now, you can see here in the same example, I tried to compute there d_i . So, you have to simply understand how are you going to write down the d_i . So, d_i is simply here $6 - 4$. Or you can also take $4 - 6$ whatever you want, because this is the difference. So, these are the values which are obtained by this $2 - 1$, I have considered here. So, this will be $6 - 20 - 14$ and these are the value of d_1, d_2, d_3, d_4 and d_5 .

(Refer Slide Time: 33:12)

Testing Hypothesis for Difference of Means (Dependent Samples): Example in R

- To test $H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$
- At 1 % level and 4 d.f., Critical Value = 4.604

$$T_c = \frac{\bar{d}}{s_d/\sqrt{n}} = \frac{-4.2}{5.67/\sqrt{5}} = -1.66$$



- Decision: Do not reject H_0
- Conclusion: There is no evidence of a significant change in the number of complaints

So, now, you simply try to find out their mean and variance and you try to compute here this statistics and exactly in the same way as you took the decision in the earlier case, you can also do the same thing over here means, if you try to take $\alpha = 0.01$, the degrees of freedom are going to be $n - 1$ that is $5 - 1$, 4 and the critical value that can be obtained from the tea table here as like this.

Now, you understand all these things. So, in this case, the value of calculated value of this data is coming down to -1.66 , calculated value is 4.604 , so you can see here this is lying somewhere here. So, the conclusion accept H_0 or do not reject H_0 . So, that means, there is no evidence of a significant change in the number of complaints after the training. So, the training was not effective.

(Refer Slide Time: 34:03)

```
Testing Hypothesis for Difference of Means (Dependent
Samples): Example in R
ya = c(6,20,3,0,4)
yb = c(4,6,2,0,0)
t.test(ya, yb, alternative = c("two.sided"), mu
= 0, paired = T)
```

So, now, I try to do the same exercise in the R software also. So, I have entered this data as ya and yb, here like this and actually this data and you can see here is given like this ya and here yb and the alternative here is two sided. $\mu = 0$, which is the value of that $H_0 \mu_1 - \mu_2 = 0$. And now, the only option here is the paired is equal to here TRUE that is T and all other options remains the same as earlier.

(Refer Slide Time: 34:39)

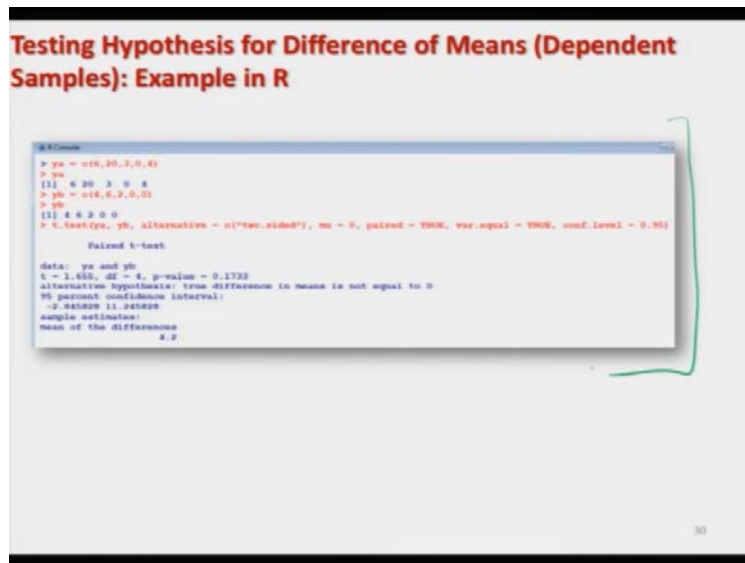
```
Testing Hypothesis for Difference of Means (Dependent
Samples): Example in R
Paired t-test
data: ya and yb
t = 1.655, df = 4, p-value = 0.1733
alternative hypothesis: true difference in
means is not equal to 0
95 percent confidence interval:
-2.845828 11.245828
sample estimates:
mean of the differences
4.2  $\bar{d}$ 
```

p > $\alpha = 0.05$

So, now, you can see here that because of the outcome and it is not difficult to understand this outcome, this is, it is telling you that this is a paired t test, the data is on ya and yb. T statistics that is $\frac{\bar{d}}{s_d/\sqrt{n}}$ that is obtained here 1.655 degrees of freedom are here 4. p-value here is 0.1733.

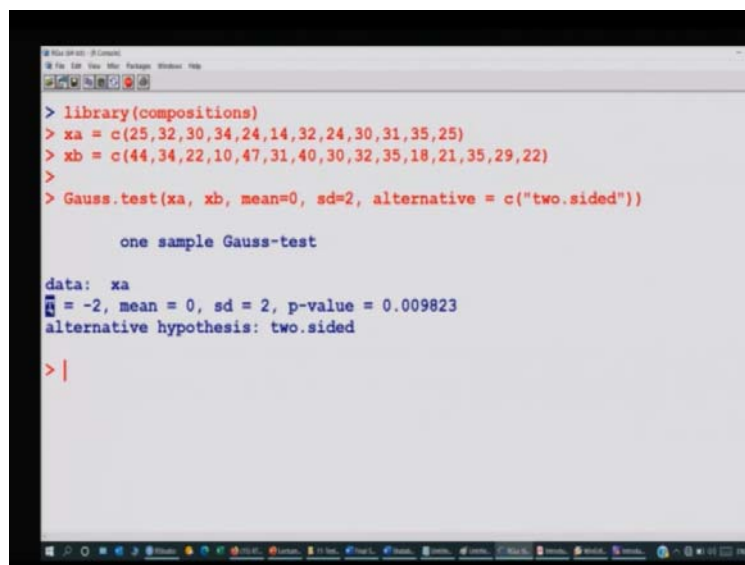
And the alternative hypothesis is that the difference in the μ is not equal to 0. that is two sided and the 95 percent confidence interval here is like this one. So, even by looking in the lower and upper bounds of the confidence interval, you can also conclude about the test of hypothesis and this is the mean of the differences that is the \bar{d} . So, you can see here that here this p-value is 0.1733. So, in this case p-value is greater than 0, say suppose $\alpha = 0.05$. So, in this case, you can see here means except the hypothesis.

(Refer Slide Time: 35:36)



And this is here, the screenshot of the same operation.

(Refer Slide Time: 35:43)



Now, let me try to show you these things on the R console. And let me try to go back in the reverse direction so that you have no problem in understanding these things. So, you can see here I am trying to take her this data on the paired t test and I am trying to copy it into R console. So, you can see here. So, firstly, let me try to load the library composition.

Well, this package is already there in my computer. So, I am only uploading it otherwise you will have to install it and if I try to give this data here x_a , x_b and this Gauss dot test, you can see here this is here the outcome you can see here. One thing you have to be careful that here it is trying to show this value as T actually this is the value of here Z. So do not get confused over these things.

So, now, we come to an end to this lecture. And now you can see that you have conducted all sorts of one sample and two sample tests for the mean. And now, I think there should not be any problem it is very easy to remember. Many times, people get confused how to remember them, if you have one sample, two sample these other two possibilities at this moment, yeah, there can be more than two samples also.

And so, in case if I am trying to consider one sample or two sample test for the mean, then there are two options, one the variance is known and variance is unknown. So, when variance is known, we are going to use the test based on the normal distribution, so called Z test, and when the variances are unknown, then we are going to use the t test. That is the only way these things are working.

And definitely once you come to the paired t test, then you have to just see that whether the sample observations in the two samples are independent or dependent and based on that you have to choose the paired t and definitely if you have more than two normal populations, like as we have considered here only two samples, but there can be more than two samples also like as somebody has collected the sample from city 1, city 2, city 3, city 4 and so on.

So, then the null hypothesis will be H_0 say μ_1 equal to μ_2 equal to μ_3 equal to μ_4 . So, whenever they are more than two means, the test of hypothesis can also be conducted for the equality of the means. And this is called as analysis of variance, which we briefly call it or popularly called as ANOVA. But I am not going to do here ANOVA but surely once you have understood this one sample and two sample test, there should not be any problem in understanding those tests for you.

So, I will now request you that you try to look into your books, try to take some problem, try to take the problem from the assignment and try to do that manually as well as on the software, so

that you can understand that whatever you have understood from the fundamental point of view, the same thing is being done in R software also. And in the next day, I will take up the last lecture of this course on the test of hypothesis for the variances. But now you are comfortable with the test of hypothesis, well I am confident about it. So, there should not be any problem in understanding the test of hypothesis for the variances. So, we try to practice and I will see you in the next lecture till then, goodbye.