Essentials of Data Science with R Software-1
Professor Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology Kanpur
Lecture 69
Test of Hypothesis for Variance in One and Two Sample

Hello friends welcome to the course essentials of data science with R software one in which we have understood the topics of probability theory and statistical inference which will lay the foundations of data science in your career. So, this lecture is the lecture where we all are going to be very happy because this is the last lecture and after that we all are free.

So, in this lecture we are just going to continue on the topic of test of hypothesis, so up to now you can see we have considered the test of hypothesis related to the mean in one sample, two sample and so on. Similarly, you can also develop the test of hypothesis for the variance in one sample and two sample.

So, the utility of such a hypothesis is exactly on the same way as we did earlier that when we are trying to consider the one sample test of hypothesis, then we are simply going to compare the value of the variance in the sample with some hypothetical value and when we are going to consider the two samples test of hypothesis for the variance, then we are going to consider the variability in the two samples with respect to some population value or their difference between the two variances in the population.

The first question comes here under what type of situations you can observe this type of thing? For example, if you ever go to a vegetable market where people are sometimes trying to sell the vegetable in the form of heaps, so they will try to prepare some heaps like as 12 oranges in every heap and then what do you do? You simply try to observe those heaps and you try to select the heap where the variation in the quality among those oranges inside the heap is not large and they are of good quality. Well there is going to be some random variation in the size and quality of the oranges because there is a natural product but you always try to make a choice with that heap where this variation is as small as possible.

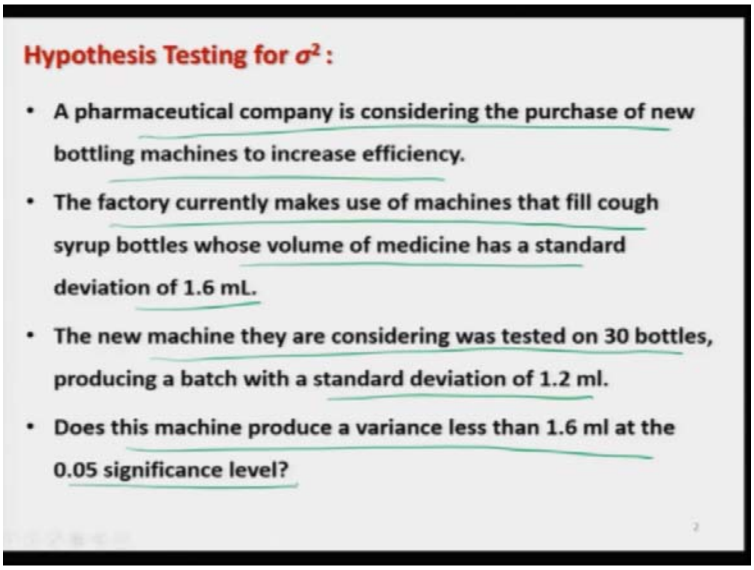
So, you always try to say I can say afford say σ^2 equal to 1. So, you will try to look for that orange where the σ^2 is going to be close to 1. So, you will simply try to observe the heap, you will try to collect the data inside your mind and you will try to compute the sample variance and you will try to compare it through some statistics with the population variance and if you feel that the variability there in the oranges is nearly the same what you expected, then you will buy the that heap. Sometime you also see when we go to the this grocery stores, suppose there is some oil in different bottles and there are say 20 bottles which are in a shelf then what you try to do?

You simply try to look at the levels of oil in the bottles and you try to choose that bottle where the level is the more. What is that? That means you are simply trying to compare the variability inside your mind with that bottle and as an owner of that shop you will always like to have those bottles where the level of the liquid is nearly the same.

So, these are the very simple situations in which you can use this test of hypothesis in data sciences, for example, if you are going to take the example of the shopping website then you can see that there will be some customer who will just come to the website, he will look for the shirts and they will buy the shirts and leave the website and there will be some customer who will be looking into the section of trousers, then shirt, then jackets and then computers and then speakers then utensils and so on so.

Then that means the behavior for the shopping for the second customer is more variation than the shopping behavior of the first type of customers. So, these types of things say you would like to know because based on that you would try to send them some offers, discount etc. So, these things are quite popular in real data application and that is why I have chosen this topic. So, let us begin the test of hypothesis on the variance but definitely this is not going to be something new for you. The steps methodologies everything that is the same the only thing what you know is that what is the statistics and what is its distribution and what is the correct amount in r to get it done. That's all. So, let us begin this lecture.

(Refer Slide Time: 05:50)



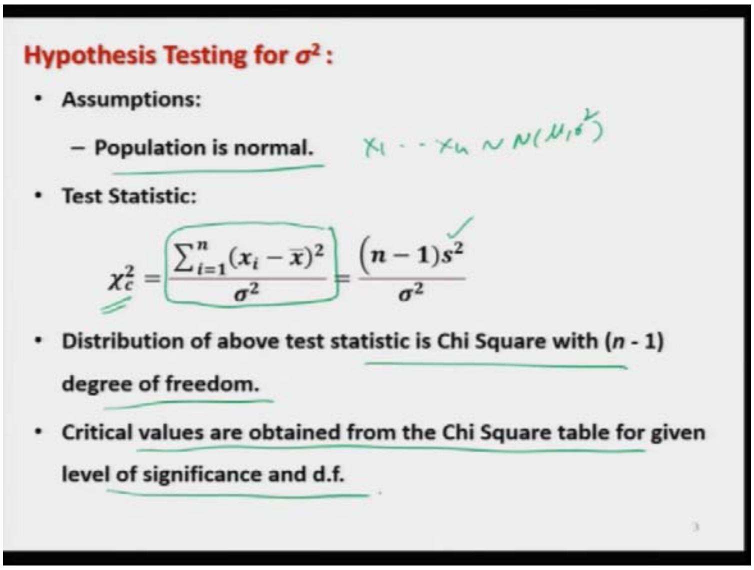
Hypothesis Testing for σ^2 :

- A pharmaceutical company is considering the purchase of new bottling machines to increase efficiency.
- The factory currently makes use of machines that fill cough syrup bottles whose volume of medicine has a standard deviation of 1.6 mL.
- The new machine they are considering was tested on 30 bottles, producing a batch with a standard deviation of 1.2 ml.
- Does this machine produce a variance less than 1.6 ml at the 0.05 significance level?

So, now we consider the test of hypothesis first in one sample case. So, there can be many examples where you would like to use such a test of hypothesis. For example, a pharmaceutical company is considering the purchase of new bottling machine to increase the efficiency, so that there is so the variation in the quantity in different bottle is not very high. The and then for example at the moment the factory is using those machines to fill the cough syrup in the bottles whose volume has a standard deviation of say 1.6 ml.

You see this random variation is always going to be there and the new machine what they are considering was tested on 30 bottles and they tried to fill up the cough syrup and then they found that the standard deviation is coming out to be 1.2 ml. So, now they have a very logical question that this new machine produces a variance less than 1.6 ml say at α is equal to 0.05 level of significance.

(Refer Slide Time: 05:52)



Hypothesis Testing for σ^2:

- **Assumptions:**
  - **Population is normal.** $X_1 \cdots X_n \sim N(\mu, \sigma^2)$
- **Test Statistic:**

$$\chi_c^2 = \boxed{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sigma^2}} = \frac{(n-1)s^2}{\sigma^2} \checkmark$$

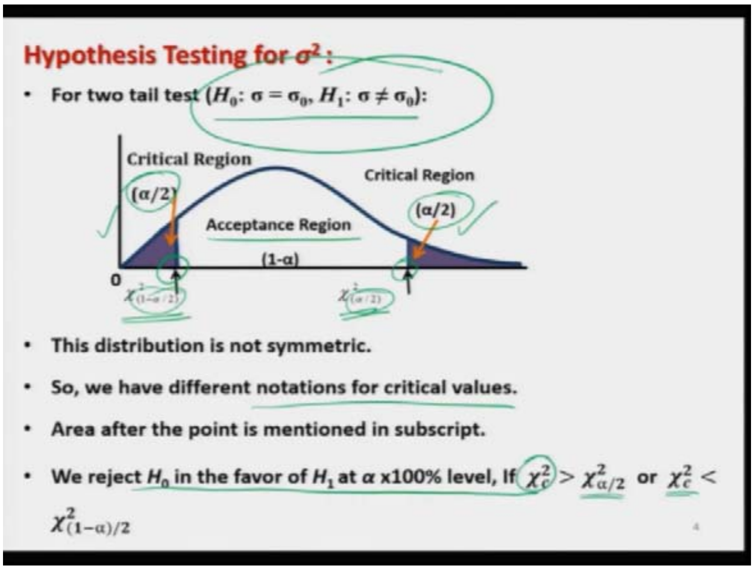
- **Distribution of above test statistic is Chi Square with (*n* - 1) degree of freedom.**
- **Critical values are obtained from the Chi Square table for given level of significance and d.f.**

So, these are the type of situations which you can see in the real life application. So, in order to develop a test of hypothesis for the variance, we simply have to assume that the population is normal, we have drawn here a sample $X_1, X_2, \ldots, X_n$ obviously from the normal population right with some mean μ and say here variance σ^2 and now the main thing is here is what is the test statistic that you want to use.

So, test static if you remember when we did the distribution of sample means and after that we also did the sampling distribution we had talked about normal distribution, Chi-square distribution, t distribution, F distribution. So, in this case in a test of hypothesis in the one sample for variance, this test is going to be based on Chi-square distribution.

So, in case if you try to take these statistics, $\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sigma^2} = \frac{(n-1)s^2}{\sigma^2}$. Then this is the test statistic right and you know that the distribution of this test statistics is a Chi-square distribution with n – 1 degrees of freedom.

(Refer Slide Time: 07:12)



Hypothesis Testing for σ^2 :

- For two tail test (H_0: $\sigma = \sigma_0$, H_1: $\sigma \neq \sigma_0$):

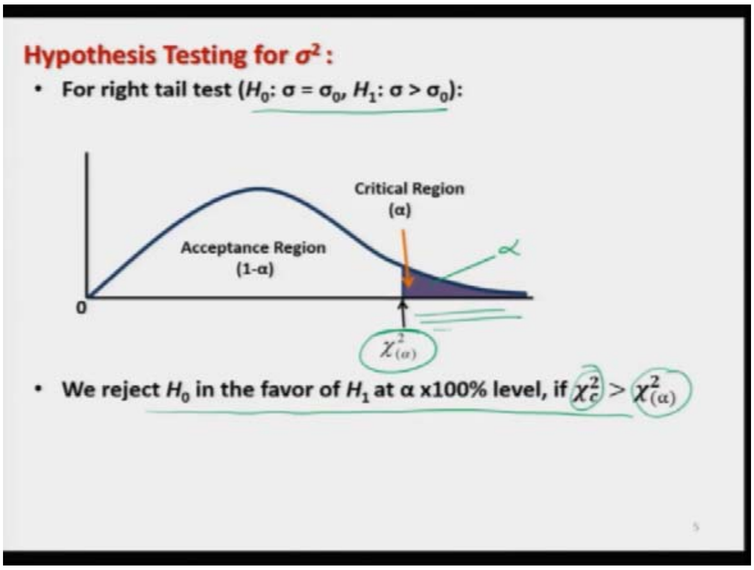
- This distribution is not symmetric.
- So, we have different notations for critical values.
- Area after the point is mentioned in subscript.
- We reject H_0 in the favor of H_1 at α x100% level, If $\chi^2_c > \chi^2_{\alpha/2}$ or $\chi^2_c < \chi^2_{(1-\alpha)/2}$

So, now and yeah if you want to obtain the critical value then obviously you know that you have to use the Chi-square table at a given level of significance and in this case suppose, if you want to conduct the test of hypothesis like H_0: $\sigma = \sigma_0$, versus H_1: $\sigma \neq \sigma_0$, then obviously the critical region is going to be on both the sides. So, this is going to be a two-sided test what you have to see here the size of the two region is going to be α/2 α/2 and the middle region which is the acceptance region is going to be 1 – α.

The main thing what you have to observe here that this Chi-square distribution is not a symmetric distribution, so the critical values which are here and here on the two sides they will be obtained as the value of the Chi-square at 1 – α/2 points and Chi-square at α/2 points. So, these values are not going to be say plus 1.96 and – 1.96 type of values.

So, that is why we have got different notation for that two critical values here right and the area what they are trying to cover that is mentioned here say α/2 or 1 – α/2 in the subscript. So, in this case also the exactly on the same length now you have done it many times I can simply say that if you are interested in this set of hypotheses which is the two sided alternative hypothesis, then we try to reject the H naught in favor of H1 at a 100 α percent level of significance if this the calculated value of the statistic is less than Chi-square 1- α/2.

(Refer Slide Time: 08:37)



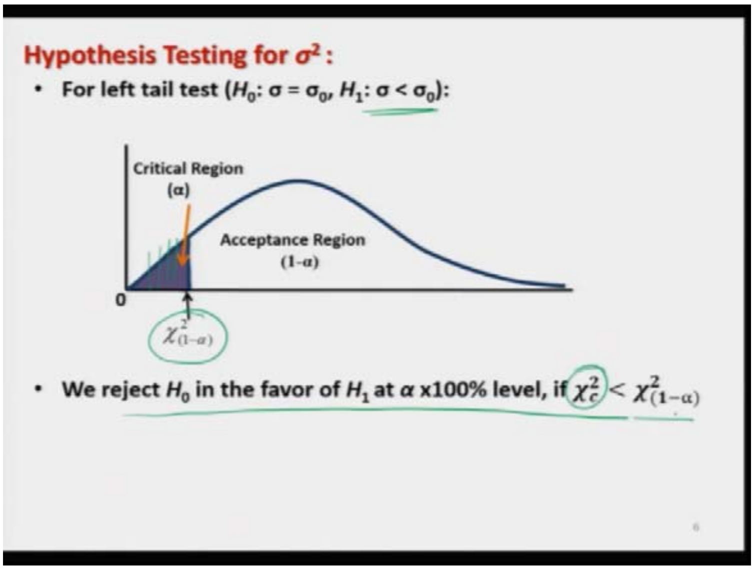
Hypothesis Testing for σ^2:

- For right tail test (H_0: $\sigma = \sigma_0$, H_1: $\sigma > \sigma_0$):

- We reject H_0 in the favor of H_1 at α x100% level, if $\chi^2_c > \chi^2_{(\alpha)}$

So, now if you want to consider the one-sided test of hypothesis say H_1: $\sigma > \sigma_0$ or H_1 is scattered $\sigma^2 > \sigma_0^2$, σ_0 is some known value that is now clear to us. So, in this case the critical region is going to be on the right hand side and this area is going to be here α and from the table you can obtain the value of Chi-square α. So, in this case we will reject the hypothesis in favor of H_1 at 100 α percent level of significance, if the calculated value of Chi-square is greater than Chi-square α, that means the value is lying somewhere here.

(Refer Slide Time: 09:18)



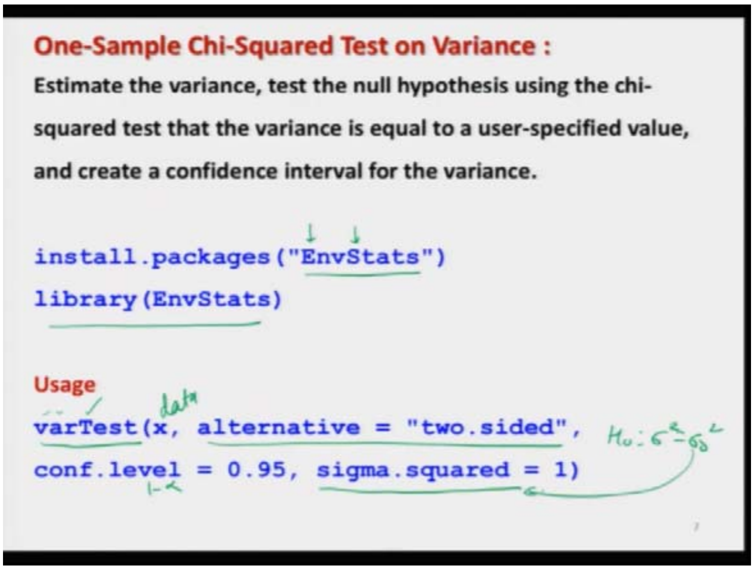
Hypothesis Testing for σ^2:

- For left tail test (H_0: $\sigma = \sigma_0$, H_1: $\sigma < \sigma_0$):

- We reject H_0 in the favor of H_1 at α x100% level, if $\chi_c^2 < \chi^2_{(1-\alpha)}$

So, that is now very simple straightforward for you and similarly if you try to take the alternative hypothesis to be σ < σ0, then obviously the critical region is going to be on the left hand side and you can obtain the value of Chi-square at 1 – α percent level of significant from the table and we reject H naught in favor of H1 at 100 α percent level of significance, if the calculated value of Chi-square is less than the tabulated value which is at 1 – α percent level of significance.

(Refer Slide Time: 09:47)



One-Sample Chi-Squared Test on Variance :

Estimate the variance, test the null hypothesis using the chi-squared test that the variance is equal to a user-specified value, and create a confidence interval for the variance.

```
install.packages("EnvStats")
library(EnvStats)
```

Usage

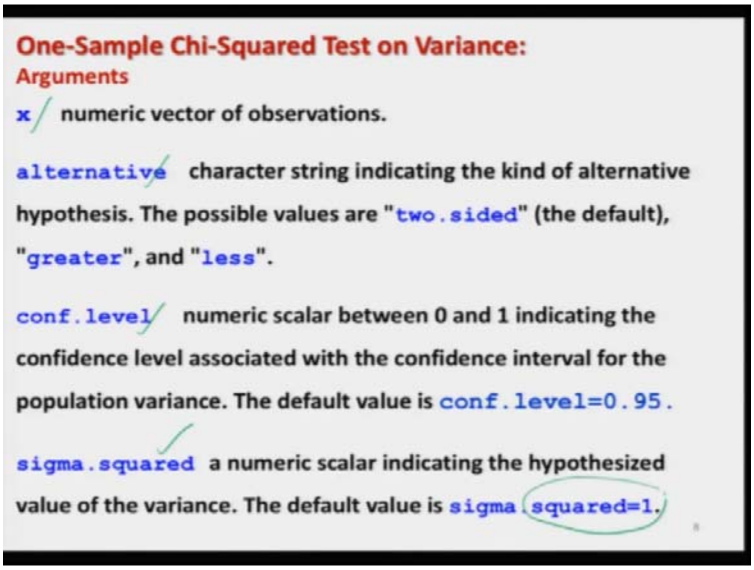
```
varTest(x, alternative = "two.sided",
conf.level = 0.95, sigma.squared = 1)
```

data
1-α
$H_0: \sigma^2 = \sigma_0^2$

Now, the main question comes here how are you going to use this in the R software? So, once again we need to install a package in the R software to conduct the test of hypothesis for these variances. So, for that we need a package here whose name is E-n-v-S-t-a-t-s, where this E and this S, they are in the upper case αbets.

So, that is from actually a book, so we try to install it and we upload it using the command library and then the command here is very simple v-a-r-T-e-s-t, this is v-a-r and T here is capital and then you have to give here the data and then you have to specify here the alternative to be two sided or less than or greater than whatever you want just exactly on the same way as we have done it earlier. Then confidence level you have to give here the value of $1 - \alpha$ and then this value of σ^2, you have to give like H_0: $\sigma^2 = \sigma_0^2$. So, this is the value of your σ_0^2.

(Refer Slide Time: 10:54)



One-Sample Chi-Squared Test on Variance:

Arguments

`x` numeric vector of observations.

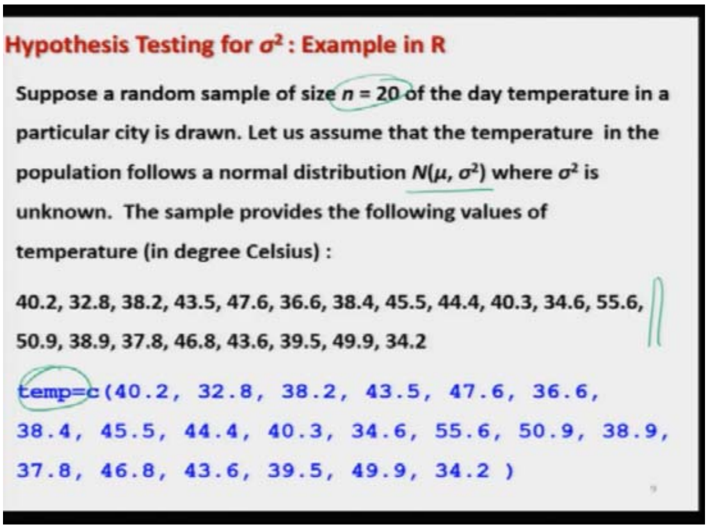
`alternative` character string indicating the kind of alternative hypothesis. The possible values are "`two.sided`" (the default), "`greater`", and "`less`".

`conf.level` numeric scalar between 0 and 1 indicating the confidence level associated with the confidence interval for the population variance. The default value is `conf.level=0.95.`

`sigma.squared` a numeric scalar indicating the hypothesized value of the variance. The default value is `sigma.squared=1.`

And now these are the same thing which I just explained you that here x is the data vector alternative is to give about the type of hypothesis that you want to consider and then confidence level that is a $1 - \alpha$ and sigma dot square is the value of σ_0^2 and you have to just keep in mind that if you do not give any value, the default value here is taken as 1.

(Refer Slide Time: 11:15)



Hypothesis Testing for σ^2 : Example in R

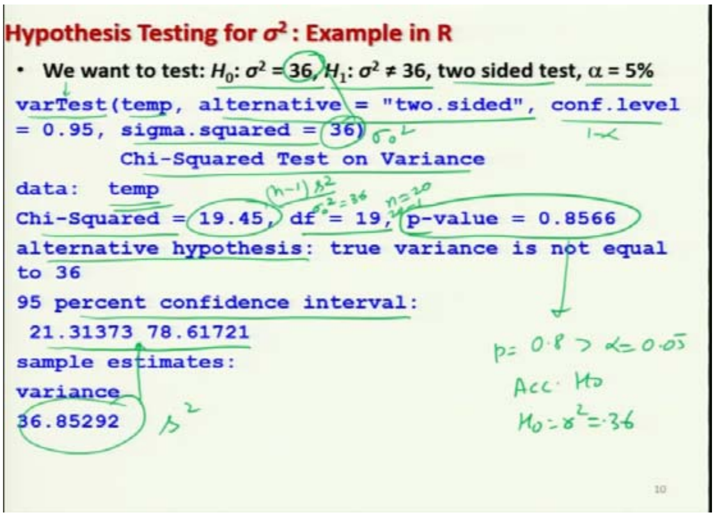
Suppose a random sample of size $n = 20$ of the day temperature in a particular city is drawn. Let us assume that the temperature in the population follows a normal distribution $N(\mu, \sigma^2)$ where σ^2 is unknown. The sample provides the following values of temperature (in degree Celsius) :

40.2, 32.8, 38.2, 43.5, 47.6, 36.6, 38.4, 45.5, 44.4, 40.3, 34.6, 55.6, 50.9, 38.9, 37.8, 46.8, 43.6, 39.5, 49.9, 34.2

```
temp=c(40.2, 32.8, 38.2, 43.5, 47.6, 36.6,
38.4, 45.5, 44.4, 40.3, 34.6, 55.6, 50.9, 38.9,
37.8, 46.8, 43.6, 39.5, 49.9, 34.2 )
```

So, now let me try to take a simple example to show you how these things can be executed. This is the same example that I considered earlier in which we have collected the day temperature on 20 different days in our city and which are assumed to follow a N(μ, σ^2) where σ^2 is unknown and this is the data and this data has been stored in the data vector t-e-m-p temperature.

(Refer Slide Time: 11:38)



Hypothesis Testing for σ^2 : Example in R

- We want to test: H_0: $\sigma^2 = 36$, H_1: $\sigma^2 \neq 36$, two sided test, $\alpha = 5\%$

```
varTest(temp, alternative = "two.sided", conf.level
= 0.95, sigma.squared = 36)
        Chi-Squared Test on Variance
data:   temp
Chi-Squared = 19.45, df = 19, p-value = 0.8566
alternative hypothesis: true variance is not equal
to 36
95 percent confidence interval:
 21.31373 78.61721
sample estimates:
variance
36.85292
```

σ_0^2

$1-\alpha$

$\frac{(n-1)s^2}{\sigma_0^2 = 36}$

$n = 20$

s^2

$p = 0.8 > \alpha = 0.05$

Acc. H_0

$H_0: \sigma^2 = 36$

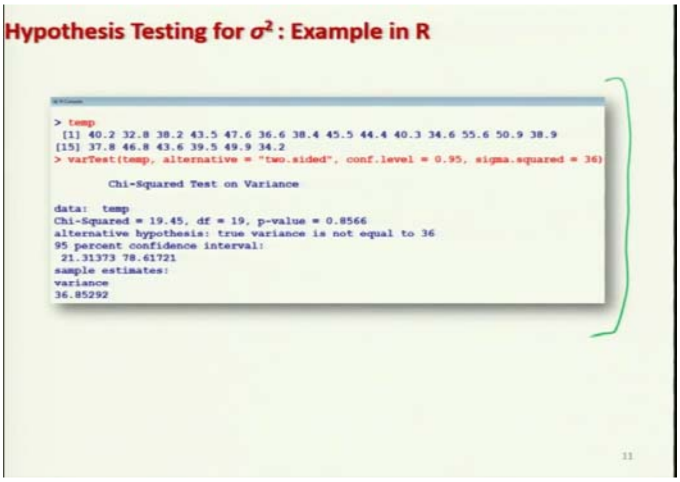
So, now let us try to consider the test of hypothesis H0: σ^2 = 36 versus H1: σ^2 ≠ 36. So, this is a two-sided test with α = 5 percent level of significance. So, now here you have to use the command here v-a-r-T-e-s-t where T is going to be in the upper-case capital letter and then you have to give here the data vector temp, alternative here is two-sided confidence level is the value of 1 – α and σ^2 =36.

So, the 36 is coming from here like this right. So, this is the value of σ_0^2. Now, if you try to execute it on the R console, you will get this type of outcome. So, first we try to understand what is this trying to explain us. So, this is the title of the test, data here is temperature, the value of Chi-square here is 19.45. So, this is the value of Chi-square statistics that $\frac{(n-1)s^2}{\sigma_0^2}$, where σ_0^2 = 36.

This degree of freedom is n – 1 and so this is and here is 20, so this becomes here 20 – 1 which is equal to 19 and p-value is calculated by the software which is 0.8566 and this is here the statement about the alternative hypothesis that you have tested here. The 95 percent confidence interval in this case will come out to be between 21.31 to 78.61 and this is here the value of s square what you have obtained on the basis of given set of data.

So, you can see here this is the value 36 which is lying inside this confidence interval 21.31 to 78.61. So, you can now see here that in this case the p-value here is 0.8, which is greater than α which α is here is 0.05 so you can say here that we are going to accept the null hypothesis. So, that means you can assume that the variance of the temperature here is 36.

(Refer Slide Time: 13:40)

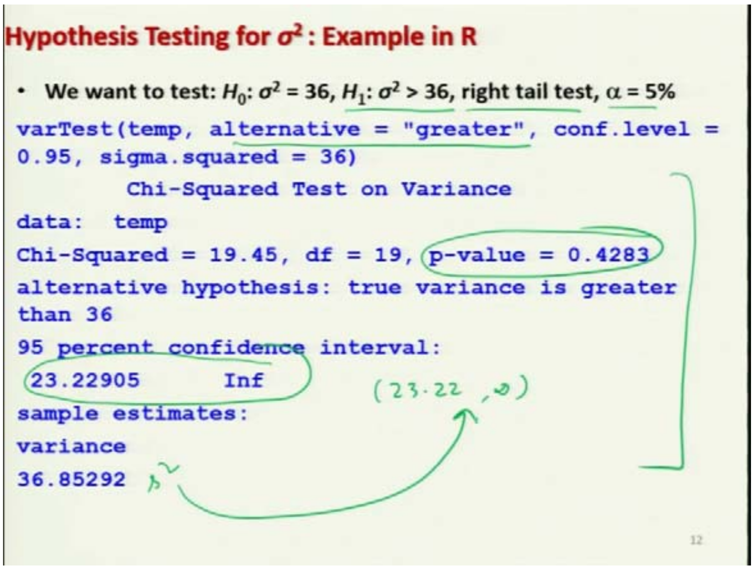


Hypothesis Testing for σ^2 : Example in R

```
> temp
 [1] 40.2 32.8 38.2 43.5 47.6 36.6 38.4 45.5 44.4 40.3 34.6 55.6 50.9 38.9
[15] 37.8 46.8 43.6 39.5 49.9 34.2
> varTest(temp, alternative = "two.sided", conf.level = 0.95, sigma.squared = 36)

        Chi-Squared Test on Variance

data:  temp
Chi-Squared = 19.45, df = 19, p-value = 0.8566
alternative hypothesis: true variance is not equal to 36
95 percent confidence interval:
 21.31373 78.61721
sample estimates:
variance
36.85292
```

Now, similarly, if you try to do it on the R console here you will get here this type of outcome. So, now you can see means if I try to copy and paste the same thing in the R console that is not contributing anything and I am sure that now you can have faith on me that the same output you will get there also. So, you can conduct it yourself just copy and paste the commands over there.

(Refer Slide Time: 14:02)



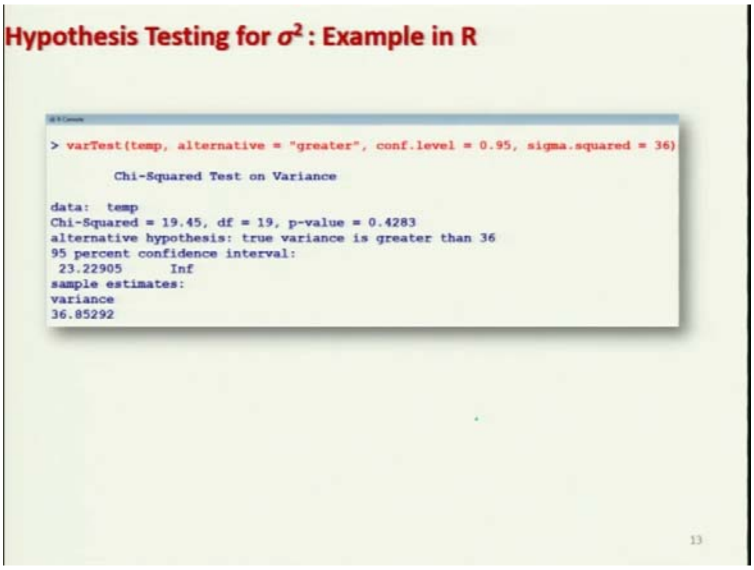
Hypothesis Testing for σ^2 : Example in R

- We want to test: H_0: $\sigma^2 = 36$, H_1: $\sigma^2 > 36$, right tail test, $\alpha = 5\%$

```
varTest(temp, alternative = "greater", conf.level =
0.95, sigma.squared = 36)
        Chi-Squared Test on Variance
data:  temp
Chi-Squared = 19.45, df = 19, p-value = 0.4283
alternative hypothesis: true variance is greater
than 36
95 percent confidence interval:
 23.22905       Inf
sample estimates:
variance
36.85292
```

Let me try to change my alternative hypothesis, it is now here H1 is now $\sigma^2 > 36$. So, this is a right tail test, α is the same. So, here in this case the all the commands remain the same, only the alternative is going to be changed to here greater and then you will get here the similar outcome. So, this is the p-value and the confidence interval here is say 23.22 to infinity. So, this is 23.22 to infinity like this one right because this is a one-sided confidence interval and you can see here the value of s square 36.85 this is also lying inside this confidence interval

(Refer Slide Time: 14:43)



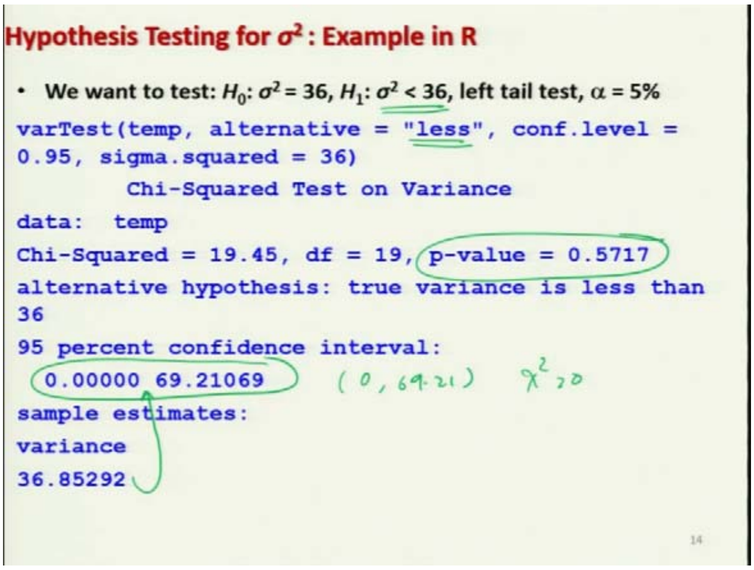
Hypothesis Testing for σ^2 : Example in R

```
> varTest(temp, alternative = "greater", conf.level = 0.95, sigma.squared = 36)

        Chi-Squared Test on Variance

data:  temp
Chi-Squared = 19.45, df = 19, p-value = 0.4283
alternative hypothesis: true variance is greater than 36
95 percent confidence interval:
 23.22905      Inf
sample estimates:
variance
36.85292
```

13

And this is the screenshot of the same operation if you try to do it on the R console.

(Refer Slide Time: 14:48)



Hypothesis Testing for σ^2 : Example in R

- We want to test: H_0: $\sigma^2 = 36$, H_1: $\sigma^2 < 36$, left tail test, $\alpha = 5\%$

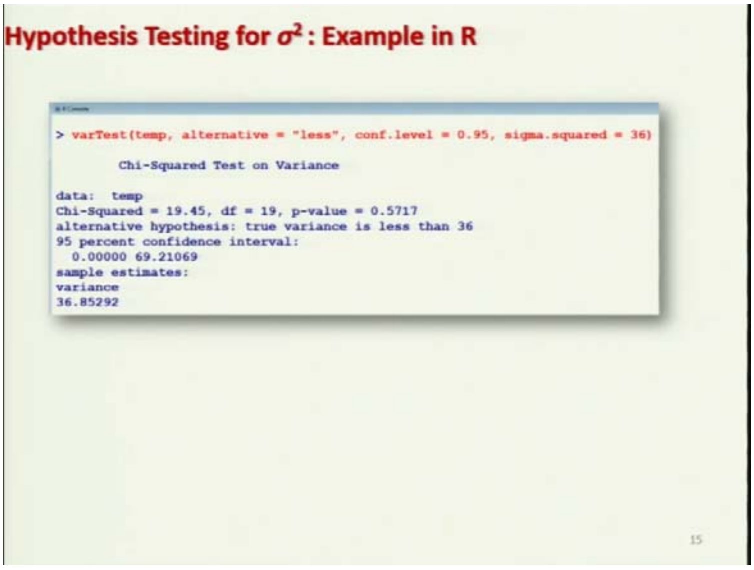
```
varTest(temp, alternative = "less", conf.level =
0.95, sigma.squared = 36)
        Chi-Squared Test on Variance
data:  temp
Chi-Squared = 19.45, df = 19, p-value = 0.5717
alternative hypothesis: true variance is less than
36
95 percent confidence interval:
 0.00000 69.21069
sample estimates:
variance
36.85292
```

$(0, 69.21)$ $\chi^2 > 0$

14

Now, once again I try to change the alternative hypothesis as σ^2 is smaller than 36 and, in this case, the earlier command remains the same, only the alternative is changed to less and you can see here the p-value is now 0.57 and the 95 percent confidence interval is starting from 0 to 69.21, so this is something like 0 to 69.21 and this is starting from 0 because you know that Chi-square variable is always taking the value which are positive greater than 0. So, that is why this is starting from 0. So, you can see here once again that the sample variant 36.85 that is also lying inside this confidence interval without any problem.

(Refer Slide Time: 15:31)

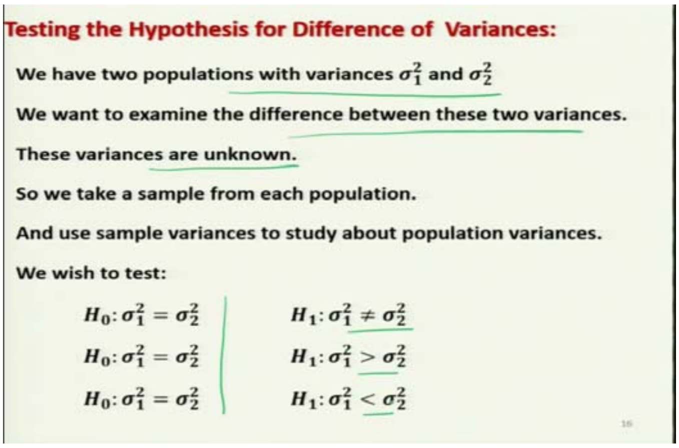


Hypothesis Testing for σ^2 : Example in R

```
> varTest(temp, alternative = "less", conf.level = 0.95, sigma.squared = 36)

        Chi-Squared Test on Variance

data:  temp
Chi-Squared = 19.45, df = 19, p-value = 0.5717
alternative hypothesis: true variance is less than 36
95 percent confidence interval:
  0.00000 69.21069
sample estimates:
variance
36.85292
```

And this is the screenshot of the same. So, I believe that now you can execute these things on the R console without any problem.

(Refer Slide Time: 15:38)



Testing the Hypothesis for Difference of Variances:

We have two populations with variances σ_1^2 and σ_2^2

We want to examine the difference between these two variances.

These variances are unknown.

So we take a sample from each population.

And use sample variances to study about population variances.

We wish to test:

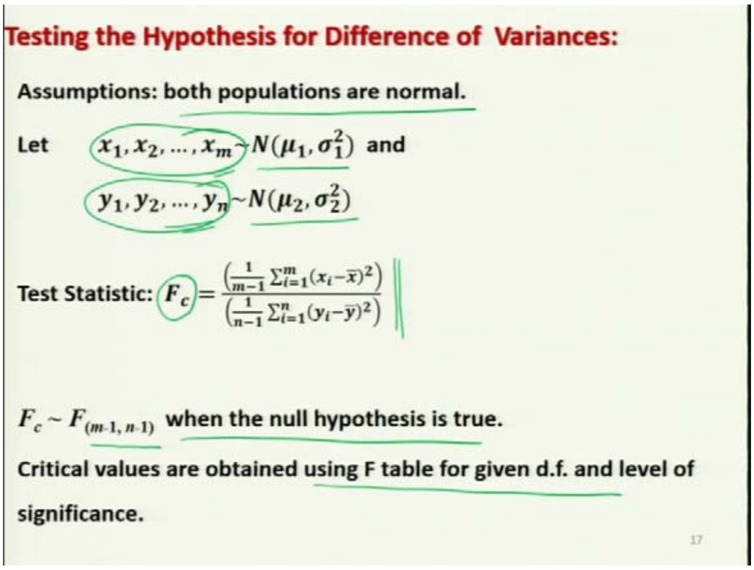
$$H_0: \sigma_1^2 = \sigma_2^2 \qquad H_1: \sigma_1^2 \neq \sigma_2^2$$

$$H_0: \sigma_1^2 = \sigma_2^2 \qquad H_1: \sigma_1^2 > \sigma_2^2$$

$$H_0: \sigma_1^2 = \sigma_2^2 \qquad H_1: \sigma_1^2 < \sigma_2^2$$

Now, after this, let me try to take the second case where we want to test the hypothesis for the difference of the variances. So, suppose we have here two population with the variance's σ_1^2 and σ_2^2 and these variances are unknown and we want to examine that the difference between the these two variances significant or not. So, what we try to do? We follow the same approach that we try to take a sample from both the populations and we try to estimate the sample variance and then we try to conduct the test of hypothesis. So, the $H_0: \sigma_1^2 = \sigma_2^2$ and $H_1: \sigma_1^2 \neq \sigma_2^2$ or $\sigma_1^2 > \sigma_2^2$ or $\sigma_1^2 < \sigma_2^2$.

(Refer Slide Time: 16:27)



Testing the Hypothesis for Difference of Variances:

Assumptions: both populations are normal.

Let $x_1, x_2, \ldots, x_m \sim N(\mu_1, \sigma_1^2)$ and

$y_1, y_2, \ldots, y_n \sim N(\mu_2, \sigma_2^2)$

Test Statistic: $F_c = \dfrac{\left(\frac{1}{m-1}\sum_{i=1}^{m}(x_i-\bar{x})^2\right)}{\left(\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2\right)}$

$F_c \sim F_{(m-1,\, n-1)}$ when the null hypothesis is true.

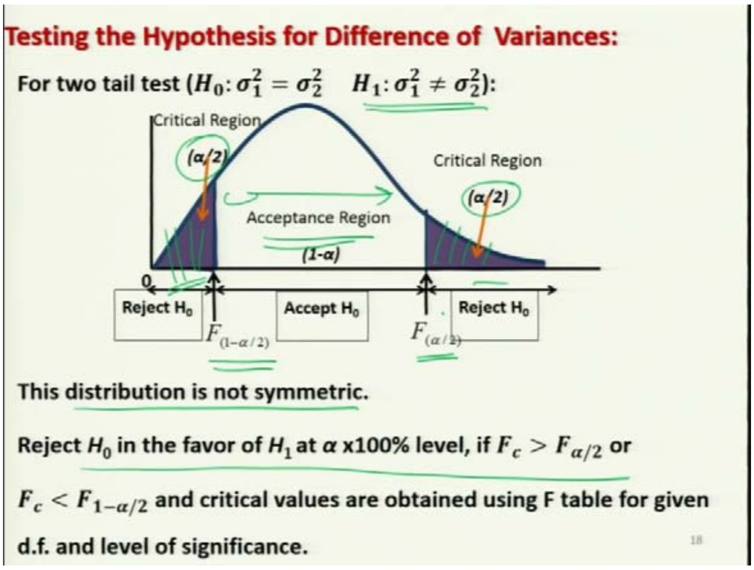
Critical values are obtained using F table for given d.f. and level of significance.

17

So, now what are we trying to do here that we are trying to make an assumption that both the populations are normal from there we try to draw here a sample of size m from $N(\mu_1 , \sigma_1^2)$ and another sample of size small n, non from $N(\mu_2, \sigma_2^2)$. So, the first sample values are indicated by $x_1, x_2, \ldots, x_m$ and the values of the second sample they are indicated by $y_1, y_2, \ldots, y_n$. Now, the next question is how to test this. Now, the next question is how to test the hypothesis and what is the relevant statistics.

So, if you try to use the Nyman Pearson lemma or this likelihood ratio test, you can find very easily that the test statistics is simply the ratio of the two sample variances. And this is statistics is denoted as a F_c because it is going to follow a F distribution with m – 1 and n – 1 degrees of freedom when H_0 is true, right and the critical values are obtained using the F table for the given degrees of freedom and level of significance. So, now you can see here you have done three distribution Chi-square, t and F, now you can see the application of those distributions.

(Refer Slide Time: 17:45)

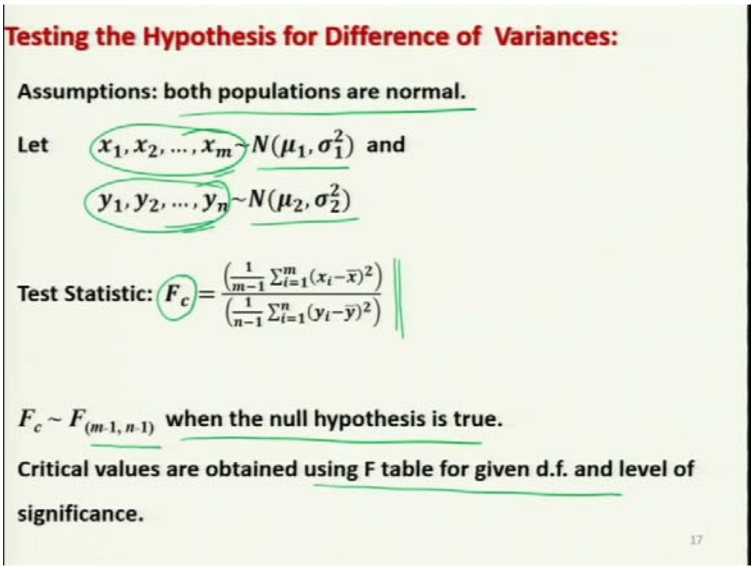


Testing the Hypothesis for Difference of Variances:

For two tail test ($H_0: \sigma_1^2 = \sigma_2^2 \quad H_1: \sigma_1^2 \neq \sigma_2^2$):

This distribution is not symmetric.

Reject H_0 in the favor of H_1 at α x100% level, if $F_c > F_{\alpha/2}$ or $F_c < F_{1-\alpha/2}$ and critical values are obtained using F table for given d.f. and level of significance.



Testing the Hypothesis for Difference of Variances:

Assumptions: both populations are normal.

Let $x_1, x_2, \dots, x_m \sim N(\mu_1, \sigma_1^2)$ and

$y_1, y_2, \dots, y_n \sim N(\mu_2, \sigma_2^2)$

Test Statistic: $F_c = \frac{\left(\frac{1}{m-1}\sum_{i=1}^{m}(x_i-\bar{x})^2\right)}{\left(\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2\right)}$

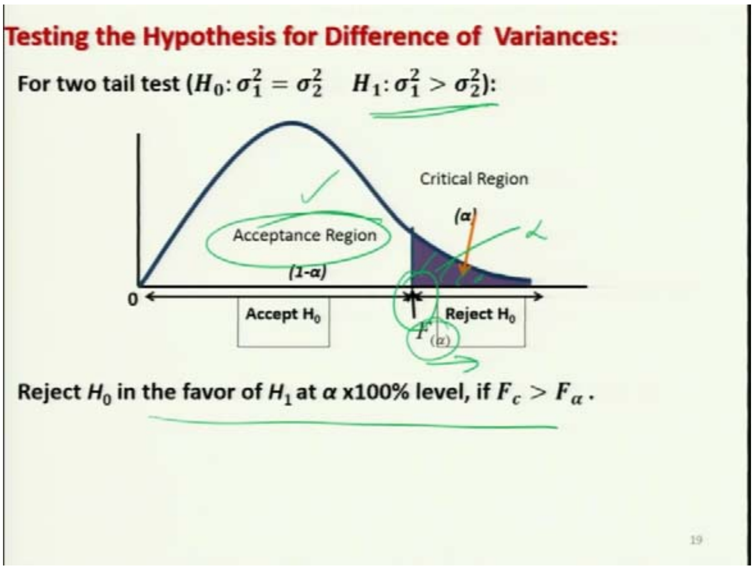
$F_c \sim F_{(m-1,\, n-1)}$ when the null hypothesis is true.

Critical values are obtained using F table for given d.f. and level of significance.

So, now in this case also, you will have the similar story that if you want to test a two-sided alternative hypothesis, then the critical reasons are going to lie on the both sides of the effort distribution. So, they are of the size α/2 and α/2 and the middle part that is x acceptance region that is 1 – α, so what you have to do is you simply have to compute the value of F on the basis of given sample data from here and you have to check whether this value is lying in the acceptance region or in the rejection region.

So, obviously if the value is greater than $F_{\alpha/2}$ and less than $F_{1-\alpha/2}$, then we are going to reject the hypothesis. And if it is a line between $F_{1-\alpha/2}$ and $F_{\alpha/2}$, then we are going to accept it and remember one thing this distribution is not symmetric, so that is why you will not hear the symmetric values of critical values on the two sides of the distribution.

(Refer Slide Time: 18:43)



Testing the Hypothesis for Difference of Variances:

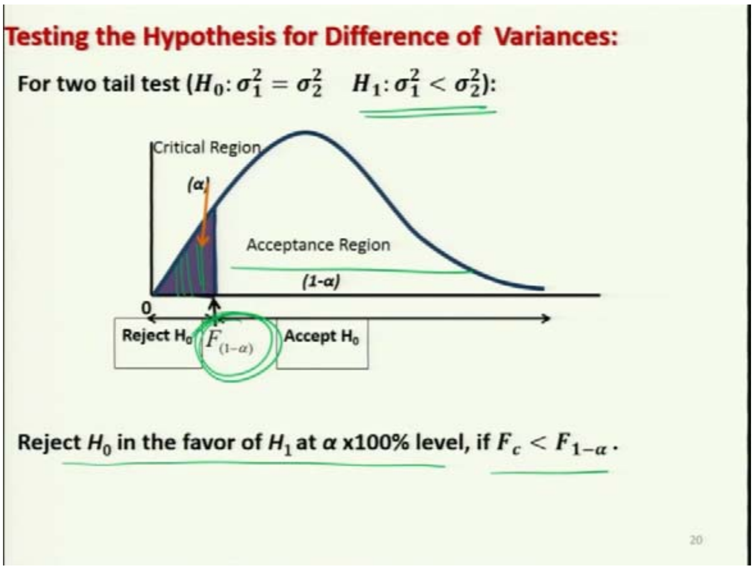
For two tail test ($H_0: \sigma_1^2 = \sigma_2^2$ $H_1: \sigma_1^2 > \sigma_2^2$):

Reject H_0 in the favor of H_1 at α x100% level, if $F_c > F_\alpha$.

19

And similarly, if you try to take here the alternative to be $\sigma_1^2 > \sigma_2^2$ then the critical region is going to lie on the right-hand side and the critical value will be somewhere here in the right-hand side, so that this area is α and so this is the region of rejection and this is here the region of acceptance in the white color. So, you are simply going to compute the value of F statistics if it lies inside the acceptance region you accept it and if this value is greater than F α, then you try to reject the null hypothesis in favor of H_1 at 100 α percent level of significance.

(Refer Slide Time: 19:20)



Testing the Hypothesis for Difference of Variances:

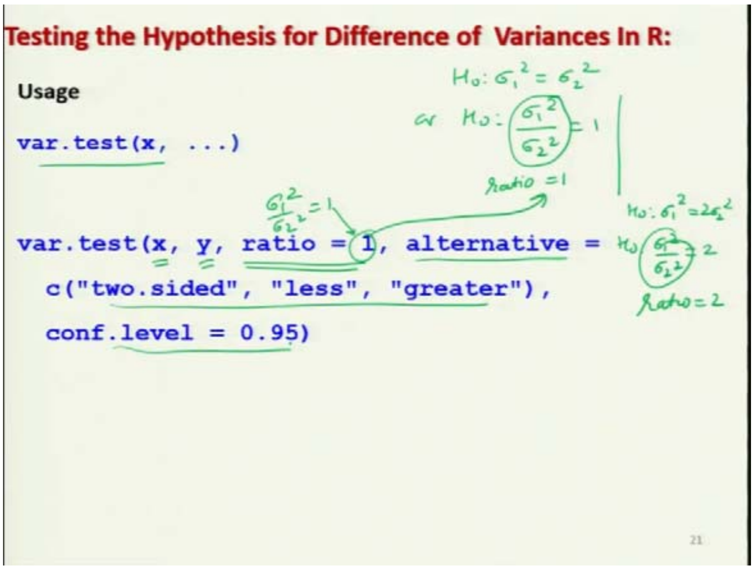
For two tail test ($H_0: \sigma_1^2 = \sigma_2^2 \quad H_1: \sigma_1^2 < \sigma_2^2$):

Reject H_0 in the favor of H_1 at α x100% level, if $F_c < F_{1-\alpha}$.

20

And similarly, if the null hypothesis is $\sigma_1^2 < \sigma_2^2$, then in that case the critical region is going to lie on the left-hand side of the f distribution and the critical value will be somewhere here. So, the acceptance region is here, so you simply try to compute the value of f on the basis of given sample of data and try to see if this value is smaller than $F_{1-\alpha}$, then you will say that reject H naught in the favor of H 1 at 100 α percent level of significance and if it is lying here somewhere in this acceptance region then you accept H_0.

(Refer Slide Time: 19:59)



Testing the Hypothesis for Difference of Variances In R:

Usage

```
var.test(x, ...)

var.test(x, y, ratio = 1, alternative =
  c("two.sided", "less", "greater"),
  conf.level = 0.95)
```

$H_0: \sigma_1^2 = \sigma_2^2$
or $H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1$
ratio =1
$\frac{\sigma_1^2}{\sigma_2^2} = 1$
$H_0: \sigma_1^2 = 2\sigma_2^2$
$H_0 \frac{\sigma_1^2}{\sigma_2^2} = 2$
ratio=2

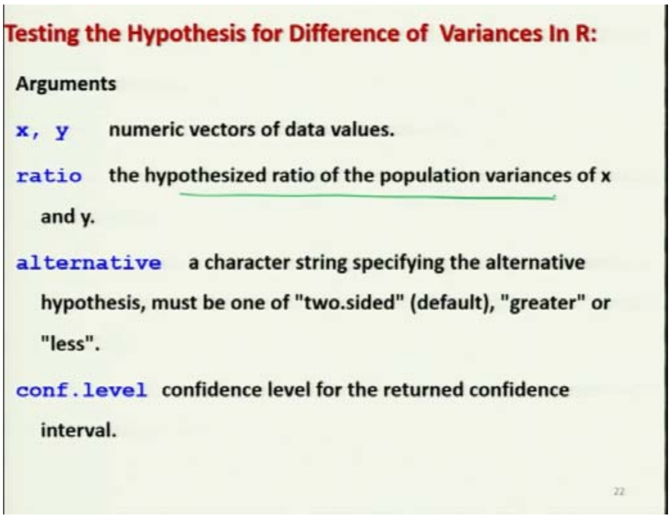
21

Now, we try to understand how this can be computed on the R software. So, one thing what you have to first understand that we are trying to test here the $H_0: \sigma_1^2 = \sigma_2^2$. So, if you try to see this can also be written as $H_0: \sigma_1^2/\sigma_2^2$=1.

So, this quantity in the R soft tier in this command is indicated by the parameter ratio. So, this means you have to give in this case ratio equal to here 1 and similarly this gives you a lot of freedom that if you want to conduct a test of hypothesis that $\sigma_1^2 = 2\sigma_2^2$, then in this case you have to write down σ_1^2/σ_2^2=2 and in this case you have to simply specify ratio equal to 2.

So, you can see here this is here the command v-a-r dot t-e-s-t and then you have to give here the data vector. So, it is so if you try to see here x is the data from the sample one, y is the data from the sample two and then we have here a parameter ratio. Ratio is here actually equal to 1 and this one is this one that is the ratio of $\sigma_1^2/\sigma_2^2 = 1$ and this one is given here and similarly other things are just like this that alternative you have to choose whatever you want two-sided less or greater and confidence level here is 0.95.

(Refer Slide Time: 21:27)



Testing the Hypothesis for Difference of Variances In R:

Arguments

`x, y` numeric vectors of data values.

`ratio` the hypothesized ratio of the population variances of x and y.

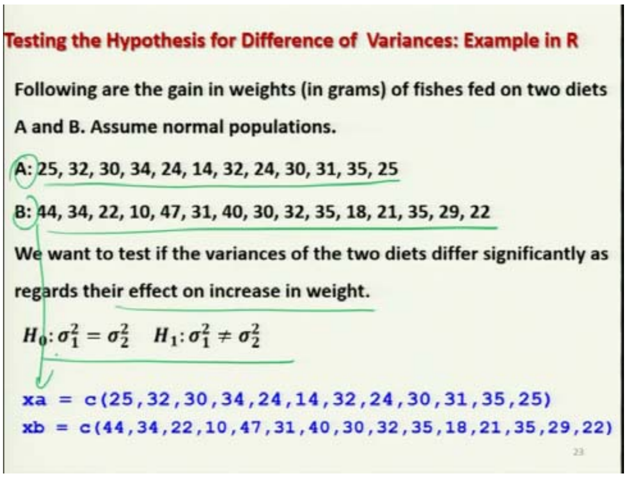
`alternative` a character string specifying the alternative hypothesis, must be one of "two.sided" (default), "greater" or "less".

`conf.level` confidence level for the returned confidence interval.

22

So, right so this is exactly what I have written here. So, you can see here the ratio here is the size ratio of the population variance of x and y.

(Refer Slide Time: 21:39)



Testing the Hypothesis for Difference of Variances: Example in R

Following are the gain in weights (in grams) of fishes fed on two diets A and B. Assume normal populations.

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

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

$H_0: \sigma_1^2 = \sigma_2^2 \quad H_1: \sigma_1^2 \neq \sigma_2^2$

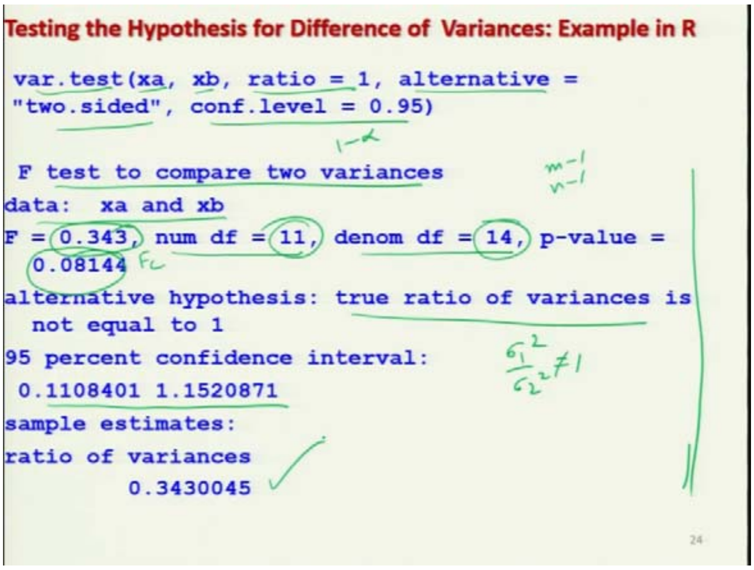
```
xa = c(25,32,30,34,24,14,32,24,30,31,35,25)
xb = c(44,34,22,10,47,31,40,30,32,35,18,21,35,29,22)
```

23

Now, let me try to take the same example that we considered earlier and try to conduct this test of hypothesis. So, in the earlier example, we had collected the weights of some fishes that means the weight they have gained after giving a diet A and after giving a diet B, assuming the normal population their increase in the weights are recorded here like this for A and for B like this in grams.

Now, we want to test if the variances of the two diets differ significantly as regard their effects on increase in the weight or not. So, we try to conduct the test of hypothesis $H_0: \sigma_1^2 = \sigma_2^2$ versus $H_1: \sigma_1^2 \neq \sigma_2^2$. So, I try to enter this data on say here A and B, here say xa and xb. So, x is my data vector, xa means the data on A and xb means the data on B.

(Refer Slide Time: 22:37)



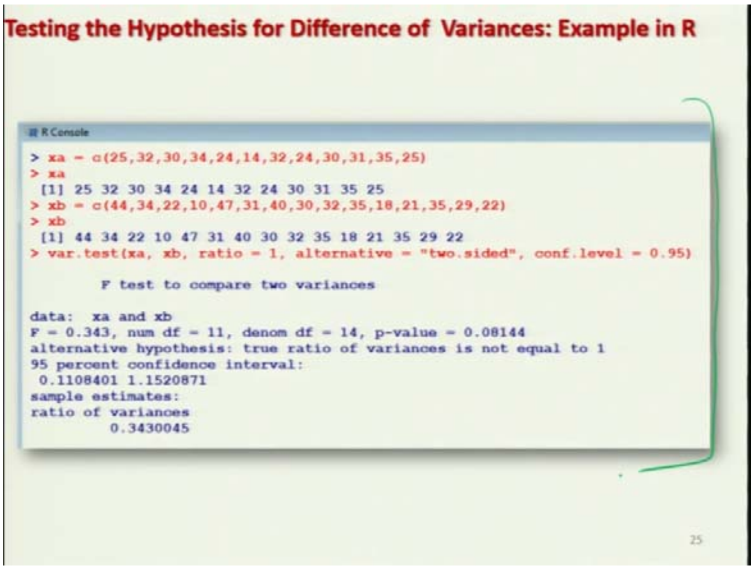
Testing the Hypothesis for Difference of Variances: Example in R

```
var.test(xa, xb, ratio = 1, alternative =
"two.sided", conf.level = 0.95)

 F test to compare two variances
data:  xa and xb
F = 0.343, num df = 11, denom df = 14, p-value =
  0.08144
alternative hypothesis: true ratio of variances is
  not equal to 1
95 percent confidence interval:
 0.1108401 1.1520871
sample estimates:
ratio of variances
         0.3430045
```

And now I try to use here the command here variance v-a-r dot t-e-s-t which is a data vector xa, xb and ratio here is 1 and alternative here is two sided and confidence level I am trying to take here 5 percent level of significance. So, you have to give here $1 - \alpha$. Now, if you try to see the outcome here, this looks like this. You can see it is not difficult to now understand for you this is the f test heading the data here is xa and xb and this is the value of f statistics that you have just computed as say F_c.

And you can see here the degrees of freedom in the numerator and degrees of freedom in the denominator they are given has to be 11 and 14 respectively. So, they are actually like $m - 1$ and $n - 1$ and p-value is coming out to be here 0.08144, alternative hypothesis is that the true ratio variance is not equal to 1, that means sigma 1 square upon sigma 2 square is not equal to 1 and 95 percent confidence interval is between 0.11 to 1.15 and the sample estimate of the ratio of variances is given to be here like this.

(Refer Slide Time: 23:47)

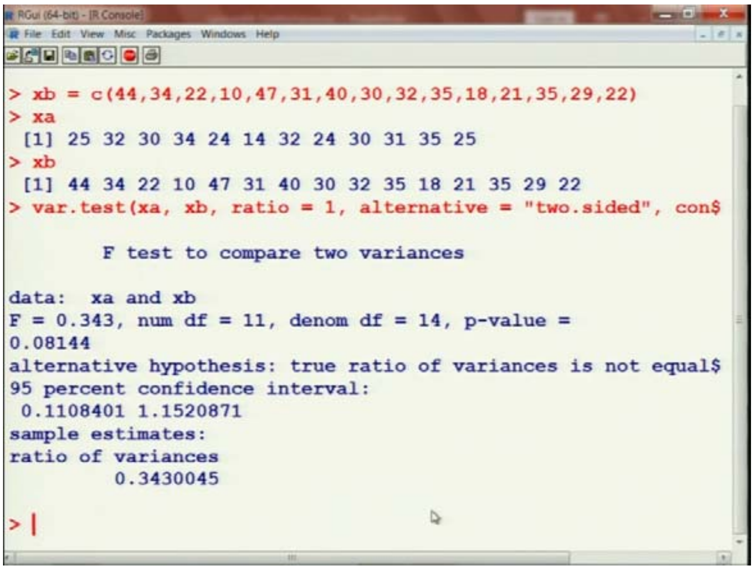


Testing the Hypothesis for Difference of Variances: Example in R

```
> 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
> var.test(xa, xb, ratio = 1, alternative = "two.sided", conf.level = 0.95)

        F test to compare two variances

data:  xa and xb
F = 0.343, num df = 11, denom df = 14, p-value = 0.08144
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1108401 1.1520871
sample estimates:
ratio of variances
         0.3430045
```

So, now you can see here and this is the screenshot of the same thing, so you can see here it is not difficult and now you know how to interpret it but let me try to show you these things on the R console.

(Refer Slide Time: 24:04)

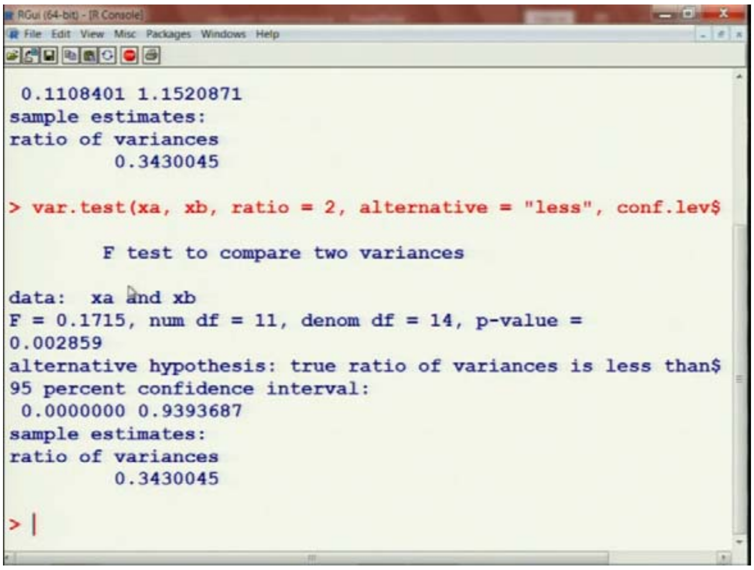


RGui (64-bit) - [R Console]
File Edit View Misc Packages Windows Help

```
> xb = c(44,34,22,10,47,31,40,30,32,35,18,21,35,29,22)
> xa
 [1] 25 32 30 34 24 14 32 24 30 31 35 25
> xb
 [1] 44 34 22 10 47 31 40 30 32 35 18 21 35 29 22
> var.test(xa, xb, ratio = 1, alternative = "two.sided", con$

        F test to compare two variances

data:  xa and xb
F = 0.343, num df = 11, denom df = 14, p-value =
0.08144
alternative hypothesis: true ratio of variances is not equal$
95 percent confidence interval:
 0.1108401 1.1520871
sample estimates:
ratio of variances
         0.3430045

> |
```

So, first I try to copy and paste this data so that I can save some time. You can see here this is the data here xa, this is the data here xb and then you have to simply use this command over here and you will get here this outcome. You can see here this is the same outcome which you have just obtained, right.

(Refer Slide Time: 24:25)


RGui (64-bit) - [R Console]
File Edit View Misc Packages Windows Help

```
 0.1108401 1.1520871
sample estimates:
ratio of variances
         0.3430045

> var.test(xa, xb, ratio = 2, alternative = "less", conf.lev$

        F test to compare two variances

data:  xa and xb
F = 0.1715, num df = 11, denom df = 14, p-value =
0.002859
alternative hypothesis: true ratio of variances is less than$
95 percent confidence interval:
 0.0000000 0.9393687
sample estimates:
ratio of variances
         0.3430045

> |
```

And even in the this case if you want to test the hypothesis like $\sigma_1^2 = 2\sigma_2^2$, so you can take the ratio equal to here 2 and you can see here now you can test this type of hypothesis and even if you want to test the hypothesis that $\sigma_1^2 < 2\sigma_2^2$ then you can change the alternative and you will get here the outcome like this. So, you can see here it is not difficult at all to conduct the test of hypothesis in this case.

Now, we come to an end to this lecture as well as in this course. Well I am not saying that I have covered all the possible topics on the test of hypothesis. There are many more tests which are available but yeah means we have a limited time over here, so that is why I have tried my best to give you some representative cases and I have tried my best to create the foundation so that once you learn these many test of hypothesis after that if you want to conduct a different type of test of hypothesis it should not be difficult for you to read them from the book and to understand them.

Well the type of test of hypothesis which I have taken here they are all finite but you can also use the last sample approximations or say asymptotic test, for example, if you want to conduct the test of hypothesis for the binomial distribution for the population proportion then you have to standardize it and that will approximately follow a normal 01 distribution. So, you have to use those things, similarly if you have a set of where you have more than 2 means or more than 2 variances, then you have different types of test for the means we have analysis of variance and for the variance, we also have a different type of test which can test the equality of the variances for more than two populations.

And even if you try to change the types of hypothesis composite versus composite, simple versus composite, composite versus simple etc. there are all sorts of combinations are there, those test of hypothesis are possible but you need to study them but you need to understand them but my simple advice is that whenever you are trying to use, you are not always trying to ah to use all of them but whatever you are trying to use them please consult a book first. A book on statistics will help you in all the topics whatever I have covered in this course. So, with these concluding remarks on this lecture I come to an end to this course also.

I hope you enjoyed the course and as I said that the topic of the course or the title of the course I have chosen as essentials of data science. I am not teaching you here data science because if you want to study the data science, you have to learn something more, you have to learn the computer programming you have to learn the computer related topics and then you have to learn many other things, good mathematics optimization techniques etc.

But definitely without having the foundation from the statistical topics you cannot learn the data science and that is what I meant when I took the title of the course to be essentials of data science but definitely I have tried my best that to avoid the mathematical proofs and derivation as µch as possible but certainly whatever I have done here, they have got a very strong mathematical theory strong mathematical proofs and those who are interested, they need to look into the book and try to understand it.

But definitely my ultimate conclusion is that well these courses can help you in understanding what you want to study, what you have to study but without reading them from the books without studying them from the standard books, proper books you cannot enhance your knowledge level. So, that is true not for this course but for all the courses actually.

So, I would suggest you that whatever topics you have done, try to read them from the books. Well I have not covered all the topics of statistic, there are many-many more right but definitely these are the topics without which you cannot proceed further. So, well I am going to end the course but your journey for the data science is going to start from today. More you study, more you understand more you use your logic that how you can employ a statistical tool in a given condition you will become a µch better data scientist what you are today, that is my promise to you and with this promise I take your leave and may god bless you all and I will see you somewhere sometime once again. Till then, good bye.