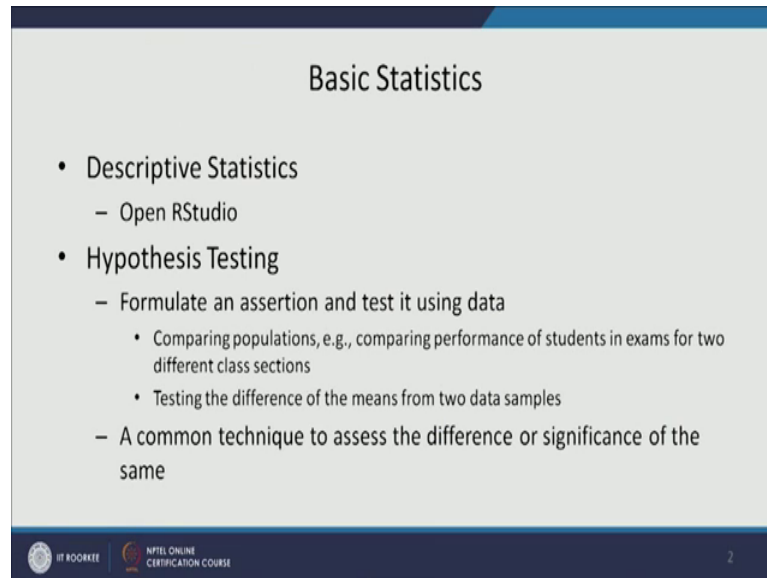


Business Analytics and Data Mining Modeling Using R
Prof. Gaurav Dixit
Department of Management Studies
Indian Institute of Technology, Roorkee

Lecture - 04
Basic Statistics Part-1

(Refer Slide Time: 00:33)



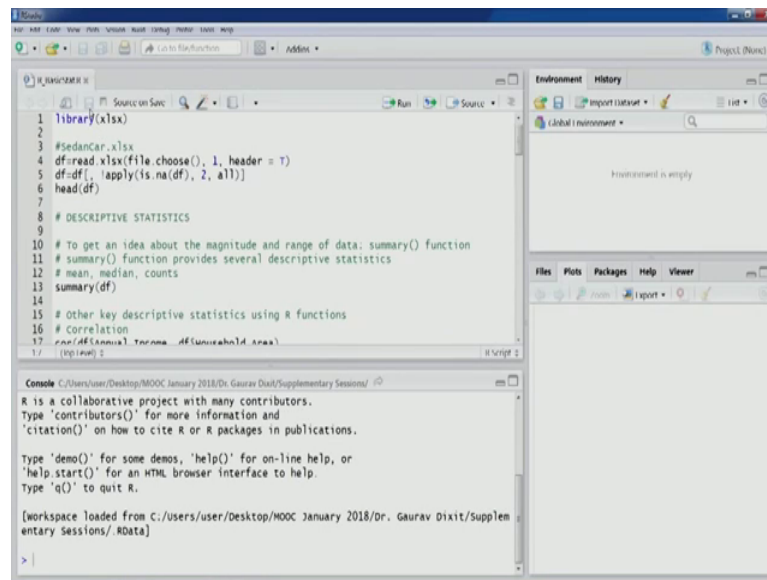
Basic Statistics

- Descriptive Statistics
 - Open RStudio
- Hypothesis Testing
 - Formulate an assertion and test it using data
 - Comparing populations, e.g., comparing performance of students in exams for two different class sections
 - Testing the difference of the means from two data samples
 - A common technique to assess the difference or significance of the same

IT ROORKEE NPTEL ONLINE CERTIFICATION COURSE 2

Welcome to the course on Business Analytics and Data Mining Modeling using R. This is our supplementary lecture number two on basic statistics using R. So, Let us start. So, as we have discussed about three types of analytics, first one being descriptive, then predictive and then prescriptive. So, we are going to cover the descriptive part, and we are going to learn some of the basic statistics using R.

(Refer Slide Time: 00:54)



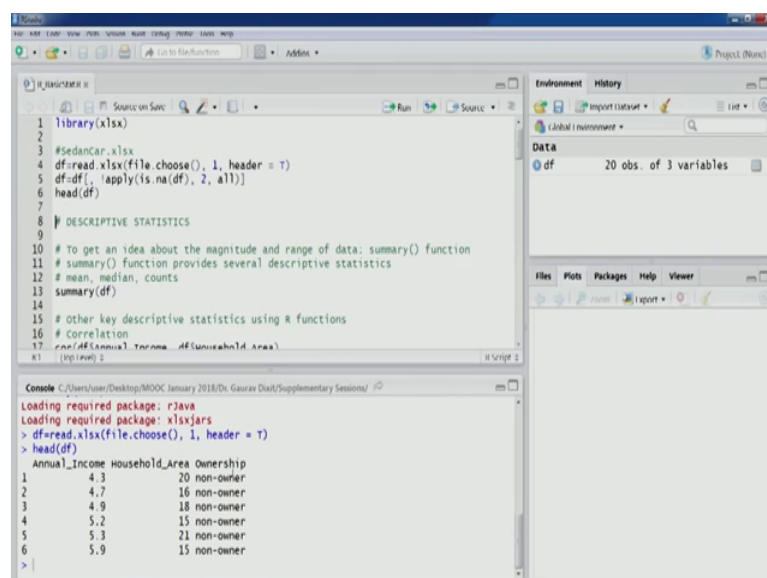
```
1 library(xlsx)
2
3 #SedanCar.xlsx
4 df=read.xlsx(file.choose(), 1, header = T)
5 df=df[, !apply(is.na(df), 2, all)]
6 head(df)
7
8 # DESCRIPTIVE STATISTICS
9
10 # To get an idea about the magnitude and range of data: summary() function
11 # summary() function provides several descriptive statistics
12 # mean, median, counts
13 summary(df)
14
15 # Other key descriptive statistics using R functions
16 # Correlation
17 cor(df$Annual_Income, df$Household_Area)
18 (topdown) 2
```

Console: C:/Users/user/Desktop/MOOC January 2018/Dr. Gaurav Dixit/Supplementary Sessions/ RData

R is a collaborative project with many contributors. Type 'contributors()' for more information and 'citation()' on how to cite R or R packages in publications. Type 'demo()' for some demos, 'help()' for on-line help, or 'help.start()' for an HTML browser interface to help. Type 'q()' to quit R.

So, Let us open RStudio. So, as we have done in the previous lecture, we are first we are required to load this particular library. Why we need this, because we want to load the data set from we want to import the data set from an excel file. Data set that we want to import is the same one the Sedan car. So, Let us execute this line. You can see in the data section this data set has been imported you can see 20 observation and 3 variables.

(Refer Slide Time: 01:41)



```
1 library(xlsx)
2
3 #SedanCar.xlsx
4 df=read.xlsx(file.choose(), 1, header = T)
5 df=df[, !apply(is.na(df), 2, all)]
6 head(df)
7
8 # DESCRIPTIVE STATISTICS
9
10 # To get an idea about the magnitude and range of data: summary() function
11 # summary() function provides several descriptive statistics
12 # mean, median, counts
13 summary(df)
14
15 # Other key descriptive statistics using R functions
16 # Correlation
17 cor(df$Annual_Income, df$Household_Area)
18 (topdown) 2
```

Environment: Data

df 20 obs. of 3 variables

Console: C:/Users/user/Desktop/MOOC January 2018/Dr. Gaurav Dixit/Supplementary Sessions/ RData

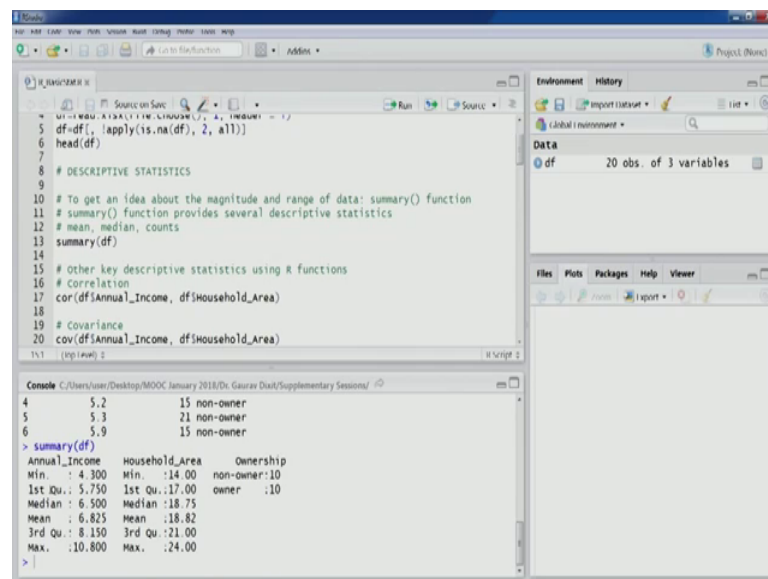
Loading required package: rJava
Loading required package: xlsxjars
> df=read.xlsx(file.choose(), 1, header = T)
> head(df)

	Annual_Income	Household_Area	Ownership
1	4.3	20	non-owner
2	4.7	16	non-owner
3	4.9	18	non-owner
4	5.2	15	non-owner
5	5.3	21	non-owner
6	5.9	15	non-owner

Again Let us have a relook at first six rows of this particular data set, you can see annual income, household area and ownership, the same variables are there. Now, let us start our

descriptive. Now, one of the first function that is popular and used quite often is summary function. Summary function in R can help you in getting the idea about the data magnitude and the range of data. Now, it also provides several descriptive statistics like mean, median and counts. So, we will see in the output. So, Let us execute this summary df. So, in df, we have three variables - annual income, household area and ownership.

(Refer Slide Time: 02:36)



```
1 df<-data.frame(Annual_Income=c(5.2, 5.3, 5.9), Household_Area=c(15, 21, 10), Ownership=c("non-owner", "non-owner", "owner"))
2 df
3 #>   Annual_Income Household_Area Ownership
4 #> 1:      5.2         15 non-owner
5 #> 2:      5.3         21 non-owner
6 #> 3:      5.9         10 owner
7
8 # DESCRIPTIVE STATISTICS
9
10 # To get an idea about the magnitude and range of data: summary() function
11 # summary() function provides several descriptive statistics
12 # mean, median, counts
13 summary(df)
14
15 # Other key descriptive statistics using R functions
16 # Correlation
17 cor(df$Annual_Income, df$Household_Area)
18
19 # Covariance
20 cov(df$Annual_Income, df$Household_Area)
```

Console Output:

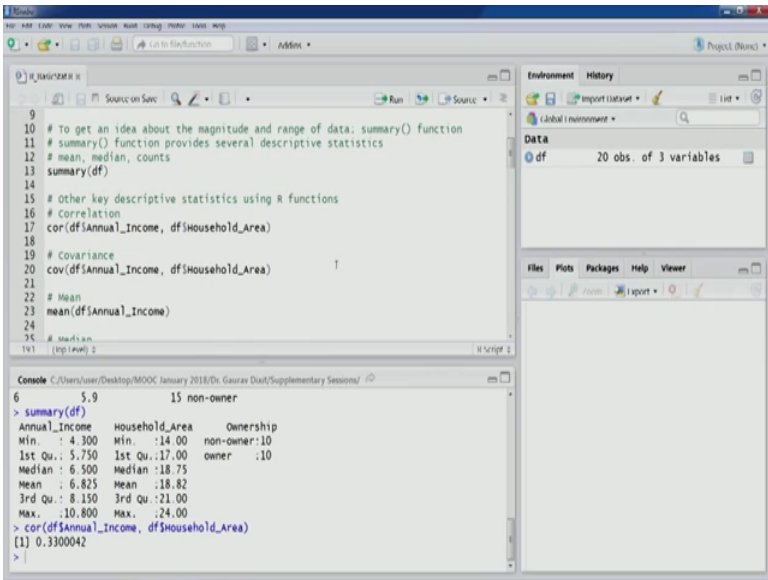
```
> summary(df)
Annual_Income  Household_Area  Ownership
min.   : 4.300   min.   :14.00   non-owner:10
1st Qu.: 5.750   1st Qu.:17.00   owner   :10
Median : 6.500   Median :18.75
Mean   : 6.825   Mean   :18.82
3rd Qu.: 8.150   3rd Qu.:21.00
Max.   :10.800   Max.   :24.00
```

We look at the output first start with let us start with annual income. You can see the values range from minimum value of 4.3 to maximum value of 10.8; mean line some were between some were at 6.8, and median line somewhere at 6.5. You can also see other things like first quartile and third quartile this is at 5.75 and 8.15. So, this quartiles also give you a you know idea about where the majority of the values are lined.

Now, let us look at the second numerical variable that is household area. So, here also you can see most of the values all the values are going to lie between minimum value of 14 and maximum value of 24. Now, majority of the values are going to lie between first quartile that is 17 and the third quartile 21, and mean lying at 18.8, and the median at 18.75. Now, these statistics are mainly for numerical variable. Now, for the categorical variable or the factor variable that we have is ownership. Now, there only the counts are displayed, some of the statistics related to numerical variable they are not applicable.

Now, Let us move on to other basic statistical methods, first one is correlations, how do we compute the correlations between two variables. So, again correlation is applicable between two numerical variable. So, we want to find out how a particular variable is correlated with another variable, so that can be done using the this functions cor function. So, we can pass on these two arguments annual income and household area and we can find out the correlation between these two variables. So, the correlation value comes out to be 0.33 for annual income and household area. So, correlation generally gives you the idea about the relationship between the variables. So, the correlation value lies between minus 1 and 1.

(Refer Slide Time: 04:52)



```
9
10 # To get an idea about the magnitude and range of data: summary() function
11 # summary() function provides several descriptive statistics
12 # mean, median, counts
13 summary(df)
14
15 # Other key descriptive statistics using R functions
16 # Correlation
17 cor(df$Annual_Income, df$Household_Area)
18
19 # Covariance
20 cov(df$Annual_Income, df$Household_Area)
21
22 # Mean
23 mean(df$Annual_Income)
24
25 # Median
26 median(df$Annual_Income)
```

Environment History

Data

df 20 obs. of 3 variables

Files Plots Packages Help Viewer

Console

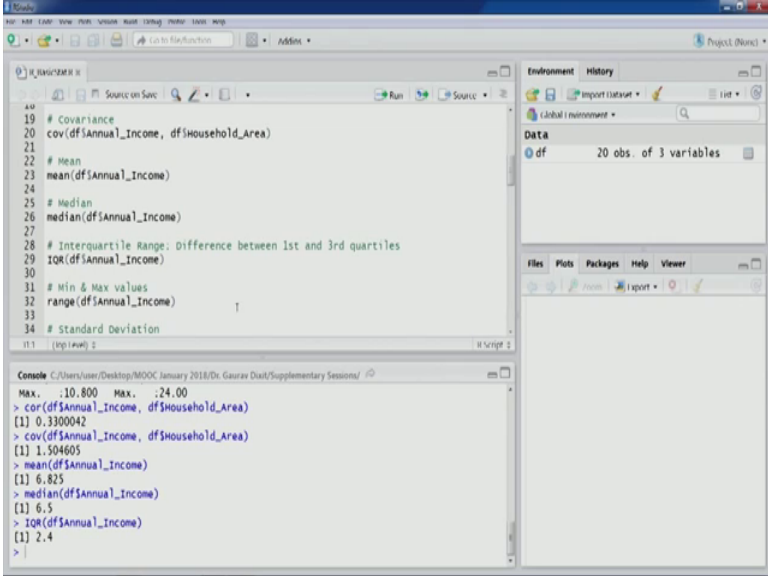
```
6 5.9 15 non-owner
> summary(df)
Annual_Income Household_Area Ownership
Min. : 4.300 Min. :14.00 non-owner:10
1st Qu.: 5.750 1st Qu.:17.00 owner :10
Median : 6.500 Median :18.75
Mean : 6.825 Mean :18.82
3rd Qu.: 8.150 3rd Qu.:21.00
Max. :10.800 Max. :24.00
> cor(df$Annual_Income, df$Household_Area)
[1] 0.3300042
>
```

Now, in this case it is plus 0.33 the value, which are closer to 1 or minus 1 signify high level of high degree of correlation and values closer to 0 signify or indicate a low level of correlation between variables. More discussion on correlation we will do in coming lectures. Now, next important-statistics is covariance. Now, covariance we have cov function in R that is available to us. So, again we can pass onto numerical variables in this case our example is about annual income and household area. Let us execute this line. Now, you can see the covariance as being computed between these two variables. Now, another covariance is again the spread of values. So, how much common spread must between these two variables is there, the overlap region that is between the these two variables can actually be indicated by covariance values.

Now, another simple statistics that we can compute using simple R functions, so mean. So, mean was something that was part of `summary` function as well, but if we are interested in just computing mean of a particular variable that can also be done using this `mean` command. So, let us execute this for annual income. You can see the value. You can values same as what is displayed in summary function. Now, similarly median also can be computed. We have this function `median` in R that can be used to compute the value.

Now, if you are interested in few more statistics, for example, inter quartile range. So, inter quartile range is the difference between first and third quartiles. As we discussed in the summary function, we get the statistics related to first quartile and third quartile, this is another way to get the same information. So, let us execute this line `iqr` function and the annual income past in as an argument and you will get the value.

(Refer Slide Time: 07:14)



The screenshot shows the R Studio environment. The script editor contains the following R code:

```
19 # Covariance
20 cov(df$Annual_Income, df$Household_Area)
21
22 # Mean
23 mean(df$Annual_Income)
24
25 # Median
26 median(df$Annual_Income)
27
28 # Interquartile Range: Difference between 1st and 3rd quartiles
29 iqr(df$Annual_Income)
30
31 # Min & Max values
32 range(df$Annual_Income)
33
34 # Standard Deviation
```

The console output shows the results of the executed commands:

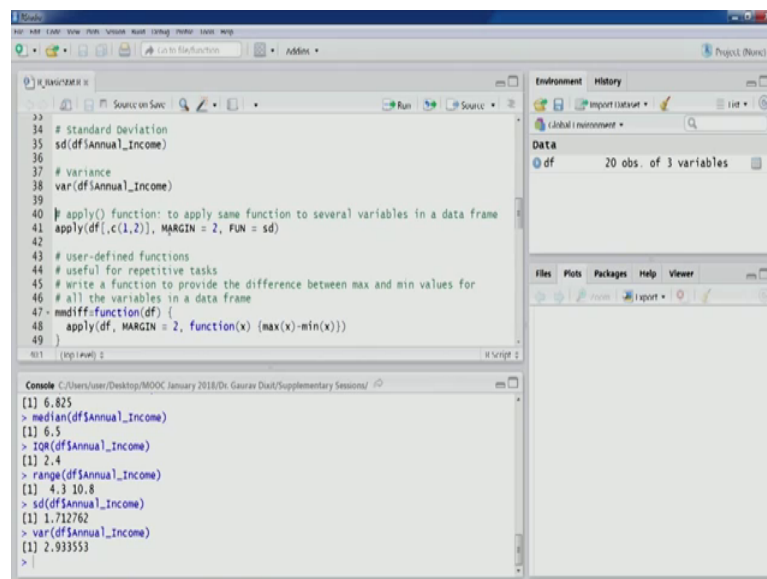
```
Max. :10.800 Max. :24.00
> cor(df$Annual_Income, df$Household_Area)
[1] 0.3300042
> cov(df$Annual_Income, df$Household_Area)
[1] 1.504605
> mean(df$Annual_Income)
[1] 6.825
> median(df$Annual_Income)
[1] 6.5
> iqr(df$Annual_Income)
[1] 2.4
>
```

The Environment pane on the right shows a data frame 'df' with 20 observations and 3 variables.

Now, if you are again in we are just interested in minimum and maximum values, so we have a direct function call `range` which can be and we can find out the minimum and maximum value. So, we do not need to depend on summary function, and we can use this standalone function which provide the specific estimate. Now, standard deviation there is this function `sd` is available in R. So, we can always compute standard deviation for any variables for annual income it counts out to be 1.7.

Similarly, if you want to compute variance of a particular variable, variance meaning the spread of values for that particular variable that can be computed using var function in R. So, you can see. So, summary function as such it covers some important some key statistics, in one go you can compute for all the variables in your data frame in your data set or you can if you are interested in one of those statistics, you can use this direct function and compute the same.

(Refer Slide Time: 08:29)



```
33  
34 # Standard Deviation  
35 sd(df$Annual_Income)  
36  
37 # variance  
38 var(df$Annual_Income)  
39  
40 # apply() function: to apply same function to several variables in a data frame  
41 apply(df[,c(1,2)], MARGIN = 2, FUN = sd)  
42  
43 # user-defined functions  
44 # useful for repetitive tasks  
45 # write a function to provide the difference between max and min values for  
46 # all the variables in a data frame  
47 mmdiff=function(df) {  
48   apply(df, MARGIN = 2, function(x) {max(x)-min(x)})  
49 }  
50  
51 (top level) 5
```

Environment History

Data

df 20 obs. of 3 variables

Files Plots Packages Help Viewer

Console

```
C:\Users\user\Desktop\MOOC January 2018\Dr. Gaurav Dutt\Supplementary Sessions/ >  
[1] 6.825  
> median(df$Annual_Income)  
[1] 6.5  
> IQR(df$Annual_Income)  
[1] 2.4  
> range(df$Annual_Income)  
[1] 4.3 10.8  
> sd(df$Annual_Income)  
[1] 1.712762  
> var(df$Annual_Income)  
[1] 2.933553  
>
```

Now, there are some important function that which are available in R which we might be required to use sometime, sometimes to transform a particular variable, sometimes do some specific task which is repetitive in nature. So, there is an already function that is available in R. So, we can use them. So, one such function is apply. So, in coming lectures, we will keep learning about many useful functions from R. So, apply function can be used. If you want to apply a function to a several variables in a data frame, this particular function can be useful. For example, we want to apply if you want to compute a standard deviation for all the numerical variables in a data frame that can be done in one go using this particular function.

So, first argument is again you need to pass on the data frame and the variables on which you want to apply. So, variables as generally you know recorded in columns, margin indicates the same thing. So, margin value up to two means that the function is to be applied column wise. Now, third argument is function f u n - fun. So, in this case, in this

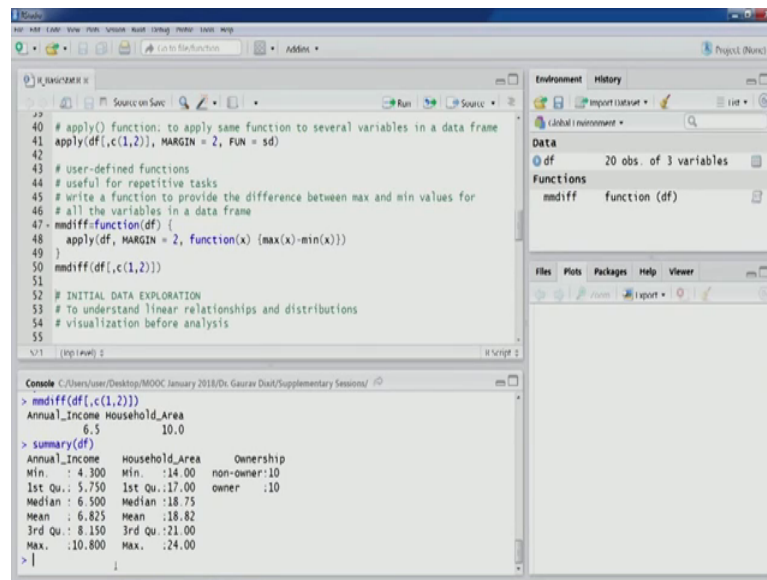
example we want to compute standard deviation values. So, sd is the that function that we have you know seen before. So, this can be passed on and we can apply. So, if you execute this line for these two variables, one variable annual income and household area, you can see the standard deviation value have been computed.

Sometimes you might have to write your own functions which are generally called user defined function. Some you know pre developed predefined function might not be available in R, and you might be required to write your own functions. So, here is an example. So, this is a very simple example, just to give you an idea about how you can write your own function and use them in your modeling R data preparation and transformation all those steps. So, this function is about providing the difference between max and minimum values for all the variables in a data frame. So, first you need to come up with the name of your function.

So, for example, because we want to compute the difference between max and min values for all the variables, so our name is mm max for min and the difference. So, mm diff is the name that I have given. And then you have to use function to define it. And then you have to mention the argument that would be allowed to pass when this case it is data frame. And then within this particular function, I have used this built in function apply this is again it again takes the first argument as data frame and this margin for column. And within this function again defining in a in a way I am again defining another function. So, this is user defined function and within apply again I am writing one more user defined functions. So, function x and max and min x. So, let us execute this code. So, that this function becomes available for us to use in future.

Now you would see in data section a functions section has being created and you would see mmdiff as the function name now this can always be called any number of time for your coding. So, mmdiff let us call this function mmdiff and we have passed these argument data frame and first and second column. So, let us execute this particular code you would see that difference between max and minimum values for these two variables annual income and household area has being computed and you can see here. If you want to verify whether your user defined function that the function written by you is working fine or not, you can do so.

(Refer Slide Time: 12:39)



The screenshot shows the RStudio IDE interface. The script editor on the left contains R code for applying a function to multiple columns of a data frame. The environment pane on the right shows the data frame 'df' with 20 observations and 3 variables, and a user-defined function 'mmdiff'. The console at the bottom shows the execution of the 'mmdiff' function on columns 1 and 2 of 'df', resulting in a summary of the differences for 'Annual_Income' and 'Household_Area'.

```
# apply() function: to apply same function to several variables in a data frame
40 apply(df[,c(1,2)], MARGIN = 2, FUN = sd)
41
42
43 # user-defined functions
44 # useful for repetitive tasks
45 # write a function to provide the difference between max and min values for
46 # all the variables in a data frame
47 mmdiff=function(df) {
48   apply(df, MARGIN = 2, function(x) {max(x)-min(x)})
49 }
50 mmdiff(df[,c(1,2)])
51
52 # INITIAL DATA EXPLORATION
53 # To understand linear relationships and distributions
54 # visualization before analysis
55
```

Environment: Global Environment

Data: df (20 obs. of 3 variables)

Functions: mmdiff (function (df))

```
> mmdiff(df[,c(1,2)])
Annual_Income Household_Area
6.5 10.0
> summary(df)
Annual_Income Household_Area Ownership
Min.: 4.300 Min.: 14.00 non-owner: 10
1st Qu.: 5.750 1st Qu.: 17.00 owner: 10
Median: 6.500 Median: 18.75
Mean: 6.825 Mean: 18.82
3rd Qu.: 8.150 3rd Qu.: 21.00
Max.: 10.800 Max.: 24.00
```

So, let us run a summary command and Let us see whether our user defined function has provided the correct output or not. So, you can see in the summary, you can see the difference between max and min value for annual income, this is 10.8 minus 4.3. So, you can see that it is 6.5 this is correct. Now, next one for household area its max value is 24 and min value is 14, difference being 10. So, you can see that. So, your user defined function is giving correct output.

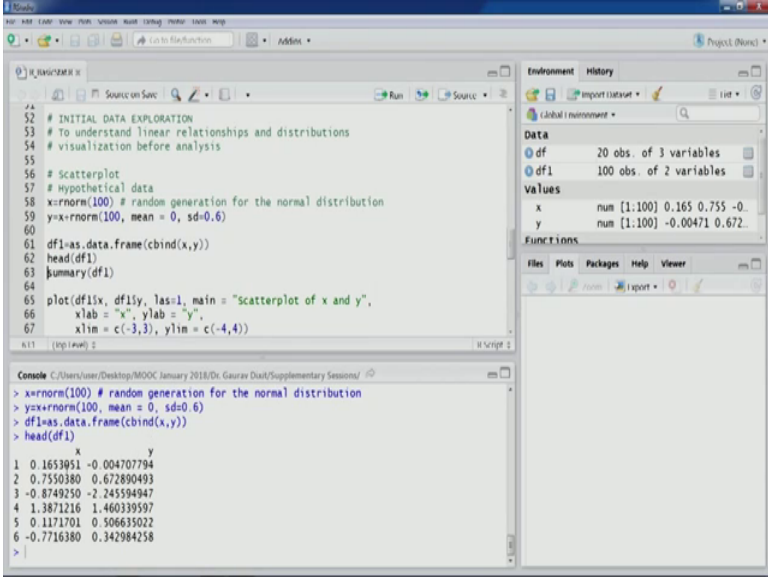
Now, let us move onto our next part that is about initial data exploration. So, whatever basic statistics that we have just discussed sometimes I mean we might require to understand a bit more for example, whether we can understand if there is any potential linear relationship between variable, whether we can understand the distribution of data. So, for that some level of visualization is required. So, now we are going to discuss some techniques related to visualization. So, these are some of the things that can be that should be done before starting the formal analysis or formal modeling.

So, one of the most important visual analysis can be done using scatterplot. So, for this, I am going to generate this hypothetical data. So, again this function R norm this can be used to generate randomly generate a data which follows normal distribution. So, R norm and the first argument that I am passing here is 100 that means, I want to generate 100 observation or 100 values. So, Let us generate values for R norm, R x. You would see in the data section x has been created, and you would see that this numeric vector have

being 100 values and the values have being generated randomly and they are also following normal distribution.

Now, we can generate another variable y. So, let us generate it like x plus R norm. Again in this case we are giving a mean specifying mean as zero standard deviation as 0.6. Let us execute this line and generate y. You can see y another vector has being created having the same number of observation 100 and the values. Now, if we want we can combine the these two variables and create a data frame, so that can be done using this hash dot data frame command. So, these two variables, they will be quartz and data frame would be created. So, let us execute this line. Now, let us see what the data looks like, so first six observation you would see that x and y you can see the these data points have been randomly generated. Let us look at the summary of this data frame. So, this is available for us.

(Refer Slide Time: 15:54)



```
# INITIAL DATA EXPLORATION
# To understand linear relationships and distributions
# visualization before analysis

# Scatterplot
# Hypothetical data
x=rnorm(100) # random generation for the normal distribution
y=x+rnorm(100, mean = 0, sd=0.6)
df1=as.data.frame(cbind(x,y))
head(df1)
summary(df1)
plot(df1[,x], df1[,y], las=1, main = "Scatterplot of x and y",
     xlab = "x", ylab = "y",
     xlim = c(-3,3), ylim = c(-4,4))
```

Environment History

DATA

- df 20 obs. of 3 variables
- df1 100 obs. of 2 variables

Values

- x num [1:100] 0.165 0.755 -0.
- y num [1:100] -0.00471 0.672.

FUNCTIONS

Files Plots Packages Help Viewer

Console

```
> x=rnorm(100) # random generation for the normal distribution
> y=x+rnorm(100, mean = 0, sd=0.6)
> df1=as.data.frame(cbind(x,y))
> head(df1)
```

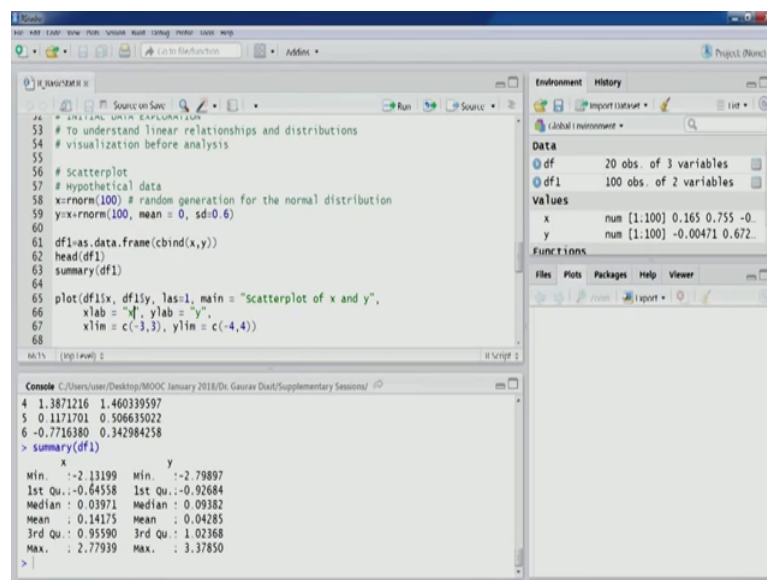
	x	y
1	0.1653951	-0.004707794
2	0.7550380	0.672890493
3	-0.8749250	-2.245594947
4	1.3871216	1.460339597
5	0.1171701	0.506635022
6	-0.7716380	0.342984258

Now, scatter plot. So, the plot command is a generic command that is available in R and can be used to generate many kinds of plots. So, in this particular case, we are trying to generate a scatter plot. So, in the plot command, we need to specify first argument should be about the variable which is going to be plotted on x-axis and then the second argument is about the variable which is going to be plotted on y-axis. And then some other las is another argument that is available mainly for the visual appeal, you can seek help to get more information on las. Then the third important argument is about main

which gives you the title of the plot. So, in this case, we have given the title of the plot as scatterplot of x and y.

Now it is important for you to level the x axis and y axis, because sometimes we are going to use data frame and the dollar notation, and then that can be taken as the default name default names for your x-axis and y-axis. So, in this case we have given the name of x-axis x and name of y-axis as y. Now, you can also specify limits for your x-axis and y-axis because sometimes your plot area might be smaller, and the it might not look good because a small portion of your plot is displaying data right. So, therefore, if you are able to restrict your limits then the plot area would you know your data points would cover majority of the plot area.

(Refer Slide Time: 17:46)



```
# LINEAR DATA EXPLORATION
# To understand linear relationships and distributions
# visualization before analysis

# Scatterplot
# Hypothetical data
x=rnorm(100) # random generation for the normal distribution
y=x+rnorm(100, mean = 0, sd=0.6)

df1=as.data.frame(cbind(x,y))
head(df1)
summary(df1)

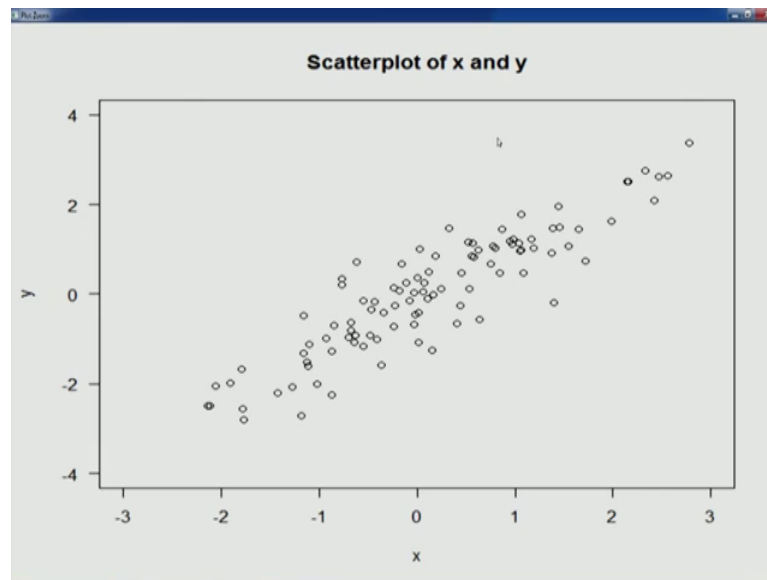
plot(df1$x, df1$y, las=1, main = "Scatterplot of x and y",
     xlab = "x", ylab = "y",
     xlim = c(-3,3), ylim = c(-4,4))
```

Console Output:

```
4 1.3871216 1.460339597
5 0.1171701 0.506635022
6 -0.7716380 0.342984258
> summary(df1)
      x              y
min.  -2.13199  min.  -2.79897
1st Qu.: -0.64558 1st Qu.: -0.92684
Median:  0.03971 Median:  0.09382
Mean   : 0.14175 Mean   : 0.04285
3rd Qu.: 0.95590 3rd Qu.: 1.02368
Max.   : 2.77939 Max.   : 3.37850
```

So, as you can see in the summary command that that we have just run the mean and max value let us look at the mean and max value, the mean value is minus 2.13 for x, and the max value is 2.77 for x. So, we can say that all the values will be going to lie you know lie within the range of minus 3 to 3, therefore we have given x limit as minus 3 to 3. Similarly, for y as well you can see that minimum value is minus 2.79 and the max value is 3.37, now all these value can lie within this range minus 4 to 4, and therefore we are given y limit as minus 4 to 4. So, let us plot this. Let us execute this code, and you would see a plot has being created.

(Refer Slide Time: 18:32)

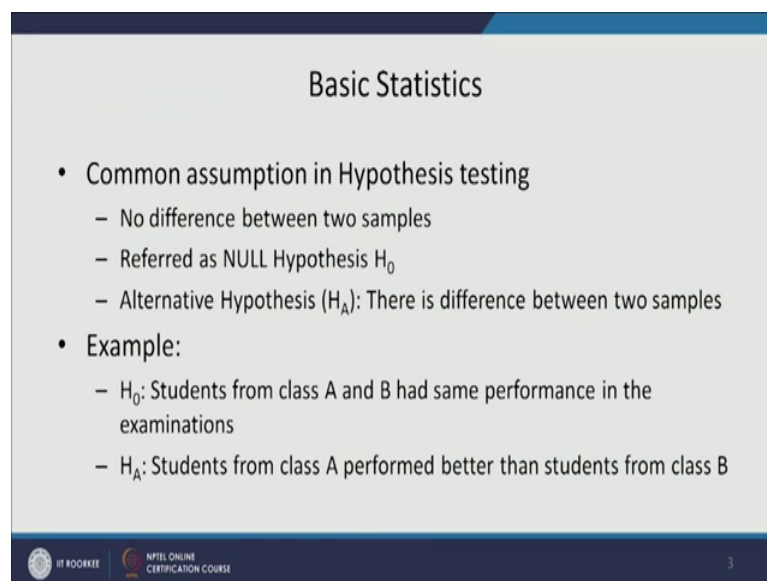


And you would see all the values and you can also form this plot, all the data points you can see, a line can be drawn from this point to this point, and it would closely fit the data. So, there seems to be linear relationship between x and y. So, why this kind of relationship is visible in this case, this is mainly because the way we have generated x and y. If you look at the way we have generated x and y, you can see x was randomly generated and then y was x plus some addition of randomly generated numbers. So, from there this linear relationship is coming.

Now, let us start our discussion on a hypothesis testing. So, hypothesis testing what hypothesis testing is about. So, this is one of the very common statistical technique that is used. So, generally whenever we are trying to formulate whenever we are trying to formulate a business problem, one part of it is going to be data mining related, analytics related or statistics related. So, therefore, mainly in when we talk about statistical modeling generally the first step is formulation of hypothesis. So, in case also we are going to learn this particular technique. So, generally it is about hypothesis testing is about comparing populations. For example, comparing performance of a students in exams for a two different class sections. So, we want to understand how class A students have performed in their exams, and whether there is significantly different from class B or is it exactly same. So, these kind of comparison could actually be performed using hypothesis testing.

So, essentially what we are doing is we are testing the difference of means from two data samples. So, one could be class A and class B and we can compare the means for these two data samples. And we can statistically we can find out whether there is difference in performance or not. So, common technique is that we use is to access the difference or significance and significance of the same. So, idea as we discussed to generally, formulate an assertion and be then test it using data.

(Refer Slide Time: 21:25)



The slide is titled "Basic Statistics" and contains the following content:

- Common assumption in Hypothesis testing
 - No difference between two samples
 - Referred as NULL Hypothesis H_0
 - Alternative Hypothesis (H_A): There is difference between two samples
- Example:
 - H_0 : Students from class A and B had same performance in the examinations
 - H_A : Students from class A performed better than students from class B

At the bottom of the slide, there are logos for "IIT ROORKEE" and "NPTEL ONLINE CERTIFICATION COURSE", and the number "3" in the bottom right corner.

Now, what are some of the common assumption in hypothesis testing. So, generally we start with that there is no difference between two samples. For example, in example that we just discussed we can assume that performance of students belonging to class A and performance of students belonging to class B is similar. So, there is no difference. So, that is the starting point for us in hypothesis testing. So, this starting point is generally referred as null hypothesis or it is denoted as H_0 . So, generally null hypothesis this is that there is no difference between two samples.

The alternative hypothesis, if we have some region to believe that the performance of class A is superior to class B or otherwise performance of class B student is superior to the same of class A, then we can say that using alternate hypothesis, which can be denoted using H_A . So, in this, we state that there is difference between two samples. Now, we are interested in knowing few more examples of hypothesis, and how we can formulate our null hypothesis, and the alternate hypothesis. So, here is one more

example. So, this one we already discussed students from class A and B that same performance in the examination being null hypothesis; and the students from class A perform better than students from class B is the alternate hypothesis.

(Refer Slide Time: 22:56)

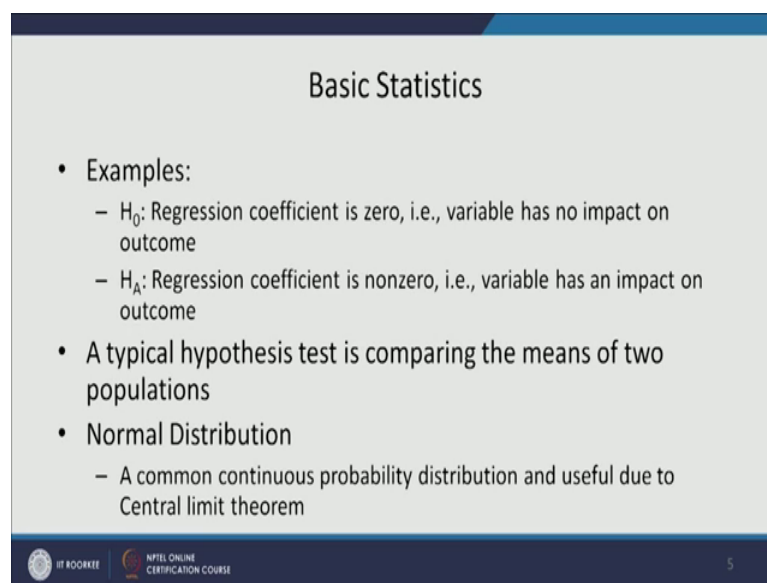
Basic Statistics

- Hypothesis test leads to:
 - Either rejection of the null hypothesis in favor of the alternative
 - Or acceptance of the null hypothesis
- Examples:
 - H_0 : New data mining model does not predict better than existing model
 - H_A : New data mining model predicts better than existing model

IT ROOKIE NPTEL ONLINE CERTIFICATION COURSE 4

Some more examples given in this particular slide, for example, new data mining model whether new data mining model does not predict better than the existing model. So, this could be null hypothesis. Alternative hypothesis could be new data mining model predicts better than existing model. So, what is going to happen after we do this hypothesis testing, so either testing results will lead to rejection of null hypothesis in favour of the alternative or acceptance of the null hypothesis.

(Refer Slide Time: 23:32)



The slide is titled "Basic Statistics" and contains the following content:

- Examples:
 - H_0 : Regression coefficient is zero, i.e., variable has no impact on outcome
 - H_A : Regression coefficient is nonzero, i.e., variable has an impact on outcome
- A typical hypothesis test is comparing the means of two populations
- Normal Distribution
 - A common continuous probability distribution and useful due to Central limit theorem

At the bottom of the slide, there are logos for "IIT ROORKEE" and "NPTEL ONLINE CERTIFICATION COURSE", and a page number "5" in the bottom right corner.

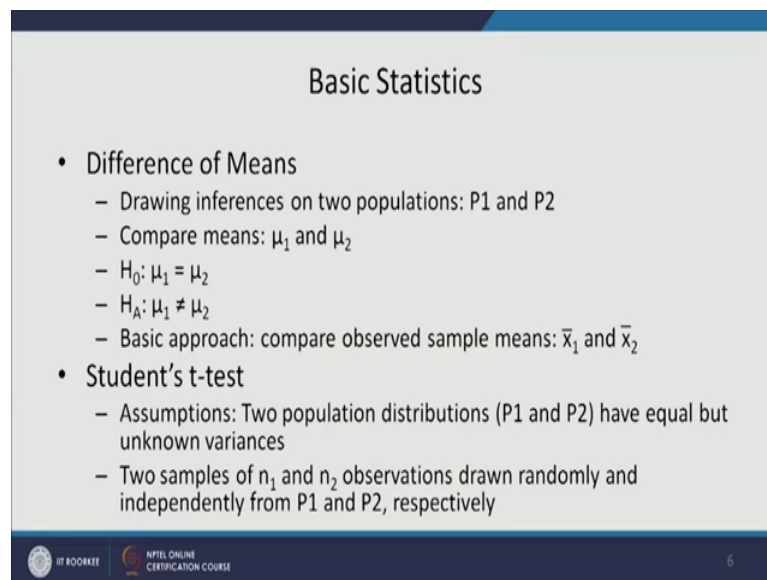
Let us look at another example. For example, this one is more related to regression analysis then we will discuss regression analysis in coming lectures, then this would seem more important information to you. So, this is important null hypothesis in a regression analysis case regression coefficient is zero, i.e., variable has no impact on outcome. The alternative hypothesis could be regression coefficient is nonzero that means, variable as an impact on outcome. So, the these are some of the examples for hypothesis formulation, and how we can state our null hypothesis and alternate hypothesis.

As we discussed before a typical hypothesis test is comparing the means of two populations. Now, at this point we need to understand another important concept. So, generally when we talk about statistical modeling and hypothesis testing in particular, generally this assumption is that the population is normally distributed. So, though the distribution is not part of this not under the scope of this particular course, but to have an understanding, normal distribution is a common is a common continuous probability distribution and useful to central limit theorem. Central limit theorem says that once we you know average of a particular sample, it tries to it follows approximately the normal distribution once the number of observation reaches 30 or more.

So, because of this whenever any population, whenever we have gather more than 30 observation, the distribution of the data, it starts following normal distribution. So,



therefore, normal distribution is a commonly occurring property, there is generally we find in different samples and populations. And therefore, it is easier for us to use this particular characteristic of distribution and then normal distribution and then use it for our hypothesis testing.

(Refer Slide Time: 25:52)



Basic Statistics

- **Difference of Means**
 - Drawing inferences on two populations: P1 and P2
 - Compare means: μ_1 and μ_2
 - $H_0: \mu_1 = \mu_2$
 - $H_A: \mu_1 \neq \mu_2$
 - Basic approach: compare observed sample means: \bar{x}_1 and \bar{x}_2
- **Student's t-test**
 - Assumptions: Two population distributions (P1 and P2) have equal but unknown variances
 - Two samples of n_1 and n_2 observations drawn randomly and independently from P1 and P2, respectively

 IIT ROORKEE  NPTEL ONLINE CERTIFICATION COURSE 6

So, generally as we discussed hypothesis testing is about difference of you know means, so we generally look for look to test difference of means. So, the idea is drawing inferences on two populations, for example, if the there are two population one is P 1 the other one is P 2. So, how we can draw inferences from these populations, so that is the main idea. So, generally this is done by comparing means. So, for example, for population one and population two mean population is μ_1 and μ_2 .

So, therefore, we can state our null hypothesis as μ_1 being equal to μ_2 , so that is null hypothesis that means both populations are same they have same mean, therefore they are same. And that the second being that mean μ_1 not equal to μ_2 that means, there is difference between these two population means, therefore between these two population. So, how do we do it because it is generally a difficult to get information about whole population. So, generally we take samples, generally we draw random samples from these populations.

So, our basic approach is to draw random samples randomly generated samples from these population and then compare observed sample means. So, we got now μ_1 and

μ_2 they are unknown, so we take sample from the population and then we will compute these observed sample means which can be denoted as \bar{x}_1 and \bar{x}_2 ; \bar{x}_1 for the population P 1, and \bar{x}_2 for the population P 2.

And then we can go about doing some hypothesis test. So, two popular hypothesis test are student's t-test and Welch t-test. So, we will go one by one. So, Let us first discuss the student's t-test. So, some of the basic assumptions that are related to student's t-test is about that two populations. So, two population distribution P 1 and P 2. So, we assume that they have equal variance. So, we do not know the variances of these two population, but we assume that them to be equal. So, only then student's t-test can actually be performed. Now, let us say we have two samples from these two samples of n_1 and n_2 observation respectively from these two populations P 1 and P 2 and they have being randomly and independently drawn from these two population. So, these are some of the assumptions related to student's t-test.


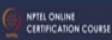
(Refer Slide Time: 28:40)

Basic Statistics

- Student's t-test
 - If P1 and P2 are normally distributed with same mean and variance
 - Then t-statistic follows a t-distribution with n_1+n_2-2 degrees of freedom

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$\text{Where } S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$$

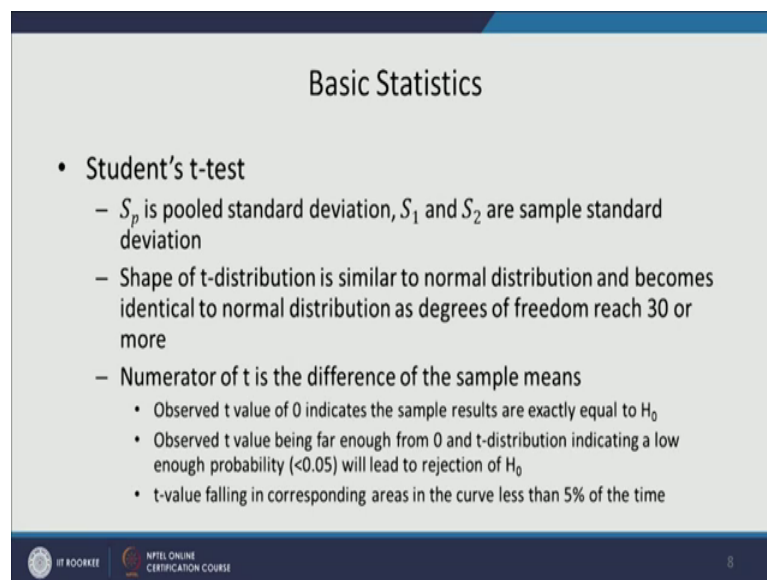


7

Now, another assumption is that is mainly about how the t-statistics is actually computed. So, if we assume that P 1 and P 2 are normally distributed this is generally the case because of the central limit theorem. So, if P 1 and P 2 are normally distributed with same mean and variance, then t-statistics follow the t-distribution in this case with n_1 plus n_2 minus 2 degrees of freedom, and this is how it can be computed. So, t-statistics can be computed as \bar{x}_1 minus \bar{x}_2 and divided by the pooled sample variance,

pool sample standard deviation and then multiplied by this particular factor square root of $\sqrt{1/n_1 + 1/n_2}$.

Now, pooled sample you know variance can be defined in this fashion. So, you can see s_1^2 is the sample variation from population one sample drawn from population one, and s_2^2 is the variance for sample 2 drawn from population 2. And you can see a kind of rated average has been taken to compute pooled sample variance. So, this is the statistics that is computed and this is under the assumption that null hypothesis is true. So, we need to understand that this has to be correct that P_1, P_2 normally distributed with same mean and variance. And we are assuming that null hypothesis is true and then we can go ahead and compute this particular t-statistics.

(Refer Slide Time: 30:27)



The slide is titled "Basic Statistics" and contains the following content:

- Student's t-test
 - S_p is pooled standard deviation, S_1 and S_2 are sample standard deviation
 - Shape of t-distribution is similar to normal distribution and becomes identical to normal distribution as degrees of freedom reach 30 or more
 - Numerator of t is the difference of the sample means
 - Observed t value of 0 indicates the sample results are exactly equal to H_0
 - Observed t value being far enough from 0 and t-distribution indicating a low enough probability (<0.05) will lead to rejection of H_0
 - t-value falling in corresponding areas in the curve less than 5% of the time

At the bottom of the slide, there are logos for "IT ROOKIE" and "NPTEL ONLINE CERTIFICATION COURSE", and a page number "8" in the bottom right corner.

So, as we said S_p is pooled standard deviation, and S_1 and S_2 are sample standard deviation. And the S_p^2 being the pooled standard variation, and S_1^2 and S_2^2 being the sample variance. Now, another point is regarding the shape of t-distribution. Now, shape of t-distribution is generally similar to normal distribution and it becomes more so when the degree of freedom reach 30 or more. As we have more and more observation into our sample then t-distribution becomes more like normal distribution. So, normal distribution and t-distribution they are also called bell curved because their shape looks like a bell.

Let us try to understand this particular t-statistics. Let us go back t , it is defined as \bar{x}_1 minus \bar{x}_2 . Now, \bar{x}_1 and \bar{x}_2 are sample \bar{x}_1 and \bar{x}_2 are observed sample means. So, if observed t values, so if \bar{x}_1 and \bar{x}_2 are quite close to each other if the observed sample means are quite close to other, the observed t values also going to be closer to 0. But it is going to be closer to 0 then the then the sample results are exactly equal to null hypothesis, therefore null hypothesis is good in that case we accepted. So, if sample observed means \bar{x}_1 and \bar{x}_2 they are close to each other or t is close to 0 then the null hypothesis is generally going to be accepted.

Now, let us understand the next point. Now, observed t value, if observed t value is far enough from 0, so it is far enough from 0, and t -distribution is indicating a low enough probability for the same then it will lead to a rejection of null hypothesis. So, if the value t that means, one of the sample one of the observed sample mean is much greater than the other one therefore a leading to a higher value of t , and the probability is also low on the lower side and that can actually lead to rejection of null hypothesis. Now, t value falling in the corresponding areas in the normal curve it should be so this also means it should be less than 5 percent of the time. So, in that case, this particular null hypothesis would be rejected. So, we will stop here, and we will continue our discussion on basic statistics using R in the next part.

Thank you.