Indian Statistical Institute

Faculty Development Program

on

Data Science

using

R

1

CONTENTS

Indian Statistical Institute

2

DESCRIPTIVE STATISTICS

DESCRIPTIVE STATISTICS

Statistics

Science of collection, analysis, interpretation and presentation of data

Analytics

Process of extracting meaningful insights by discovering patterns and relationships in the data

Indian Statistical Institute

DESCRIPTIVE STATISTICS

Data set

Number of tasks completed per hour

Productivity	
69	71
70	68
72	70
71	67
70	69
68	72
73	70
69	71

5

Indian Statistical Institute

DESCRIPTIVE STATISTICS

Data

The values or figures assigned to a metric

Can be numeric or non numeric

Example

Metric: Productivity

Data: 70, 72, 69, etc

6

Indian Statistical Institute

DESCRIPTIVE STATISTICS

Use of Statistics

To know what happened in the past

To know approximately what will happen in future if things remain more or less the same

Example

Approximately how many tasks will be completed in the next hour?

Is it same as the last value as the last value is close to the future?

The difference between the last value and previous value is only 1 (71 – 70).So adding 1 to last value will give the next future value?

Approximately how many tasks will be completed two hours down the line?

Is mean, median, etc give better estimate of future value? Why?

7

Indian Statistical Institute

DESCRIPTIVE STATISTICS

Use of Statistics

Generally the values in a data set will not be same

Values will be not only changing but change without any particular trend or pattern

Example

If we collect another 16 hours productivity data, the value may not be same as 69, second value as 70, etc?

Similarly the difference between the 1st and 2nd value or 2nd and 3rd value , etc may not be same.

It is possible that none of the values will be repeated in the new dataset

How will you make projections about future?

Explore what is remaining more or less consistent even if the values are changing.

Use it for future projections, generalizations, etc

8

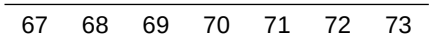
Indian Statistical Institute

DESCRIPTIVE STATISTICS

Productivity Data

69	71
70	68
72	70
71	67
70	69
68	72
73	70
69	71

Demonstration



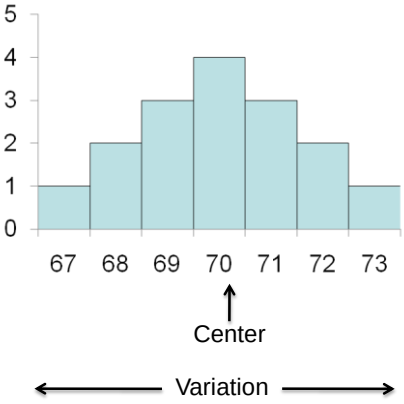
Indian Statistical Institute

DESCRIPTIVE STATISTICS

Interpretation

- Only thing remain more or less consistent is the shape
- All projections or generalizations need to be made using shape
- Shape represents the distribution
- The properties of the shape also will remain more or less constant
- The properties can be used to make projections

Productivity



Indian Statistical Institute

DESCRIPTIVE STATISTICS

Measures of Central Tendency

Mean or Average

69	71
70	68
72	70
71	67
70	69
68	72
73	70
69	71

Computation & Interpretation

Centre of gravity of the data

Sum of all values divided by the number of values

Mean = (69 + 70 + - - - + 71) / 16
= 70

On an average 70 tasks are completed in an hour

If we collect data on productivity for another 16 or 20 hours and calculate the mean it will be equal or very close to 70.

Indian Statistical Institute

DESCRIPTIVE STATISTICS

Measures of Central Tendency

Median

69	71
70	68
72	70
71	67
70	69
68	72
73	70
69	71

Computation & Interpretation

Value dividing the data set into two equal parts

After arranging the data in ascending or descending order, the value in the middle

Suppose there are n values:

If n is odd, median is the value in the $(n + 1)/2^{th}$ position

If n is even, median is the average of $n/2^{th}$ and $(n + 2)/2^{th}$ observation

Indian Statistical Institute

DESCRIPTIVE STATISTICS

Measures of Central Tendency

Median

Position	Value	Position	Value
1	67	9	70
2	68	10	70
3	68	11	71
4	69	12	71
5	69	13	71
6	69	14	72
7	70	15	72
8	70	16	73

Computation & Interpretation

n = 16

n/2 = 8

(n+2)/2 = 9

Median = Average of 8th and 9th position values

Median = (70 + 70)/2 = 70

Half of the hours the productivity will be less than 70 tasks and remaining 50% of hours productivity will be higher than 70

Indian Statistical Institute

DESCRIPTIVE STATISTICS

Variable Data Summarization: Measure of Variation or Spread

Range: Definition

Range: Maximum value – Minimum Value

Example:

5	4	7	3	2
15	9	8	5	2

Maximum Value = 15

Minimum Value = 2

Range = 15 – 2 = 13

Indian Statistical Institute

DESCRIPTIVE STATISTICS

Variable Data Summarization: Measure of Variation or Spread

Range: Issues

It depends only on extreme values

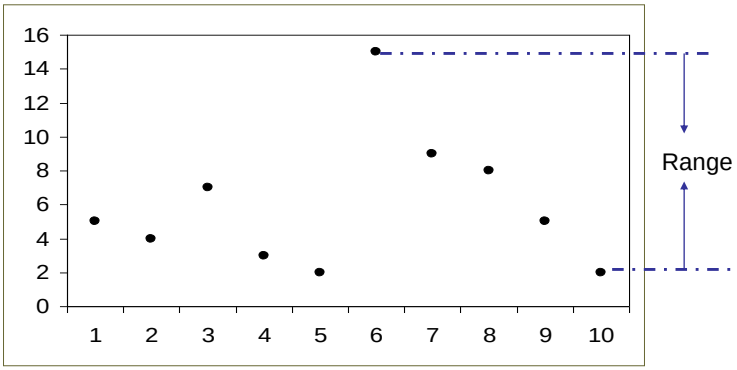
Hence affected by outliers

Indian Statistical Institute

DESCRIPTIVE STATISTICS

Variable Data Summarization: Measure of Variation or Spread

Range: Issues



Indian Statistical Institute

DESCRIPTIVE STATISTICS

Variable Data Summarization: Measure of Variation or Spread

Standard Deviation: Example:

5	4	7	3	2
15	9	8	5	2

Step 1:

Calculate Mean

Mean = 6

Indian Statistical Institute

DESCRIPTIVE STATISTICS

Variable Data Summarization: Measure of Variation or Spread

Standard Deviation: Example:

5	4	7	3	2
15	9	8	5	2

Step 2:

Take deviations from Mean

-1	-2	1	-3	-4
9	3	2	-1	-4

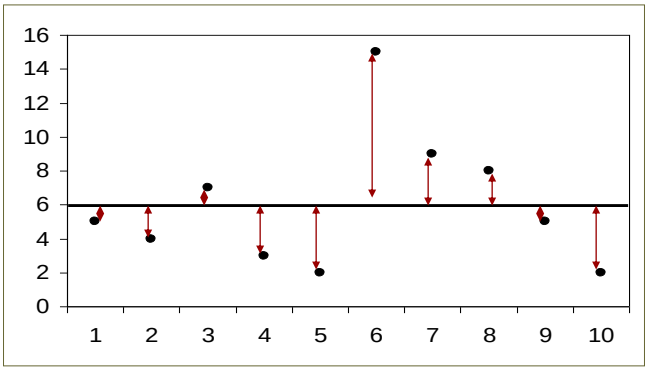
DESCRIPTIVE STATISTICS

Variable Data Summarization: Measure of Variation or Spread

Standard Deviation: Example:

Step 2:

Take deviations from Mean



DESCRIPTIVE STATISTICS

Variable Data Summarization: Measure of Variation or Spread

Standard Deviation: Example:

Step 3:

Since some values are positive & rest are negative, while taking sum they will cancel out.

So square the values & Sum

1	4	1	9	16
81	9	4	1	16

Sum = 142

Indian Statistical Institute

DESCRIPTIVE STATISTICS

Variable Data Summarization: Measure of Variation or Spread

Standard Deviation: Example:

Step 4:

Standard Deviation = √(Sum of Squares / (n -1))

= √(142 / (10 -1))

= √ 15.77 = 3.972

MS Excel Function

Standard Deviation = stdev(range of values)

Indian Statistical Institute

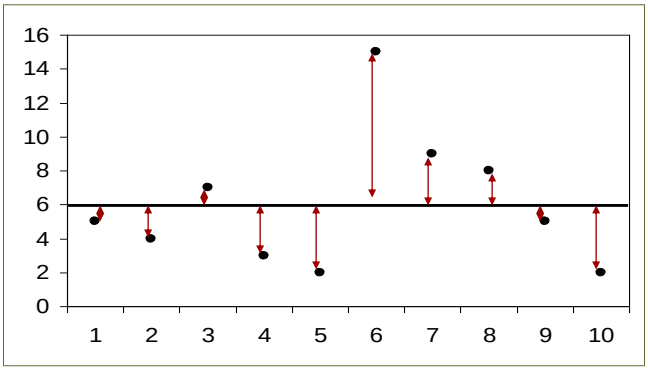
DESCRIPTIVE STATISTICS

Variable Data Summarization: Measure of Variation or Spread

Standard Deviation: Definition

Square root of the average squared deviation from mean

Indicates On an average how much each value is away from the Mean



Indian Statistical Institute

DESCRIPTIVE STATISTICS

Measures of Variation

Range

69	71
70	68
72	70
71	67
70	69
68	72
73	70
69	71

Computation & Interpretation

Difference between the maximum and minimum value of the data

Maximum = 73

Minimum = 67

Range = 73 – 67 = 6

Indian Statistical Institute

DESCRIPTIVE STATISTICS

Measures of Variation

Standard Deviation

69	71
70	68
72	70
71	67
70	69
68	72
73	70
69	71

Computation & Interpretation

Square root of the average of the square of the deviations from the mean

On an average how much each values are distributed around the center

Step 1

Compute mean = 70

Indian Statistical Institute

DESCRIPTIVE STATISTICS

Measures of Variation

Standard Deviation

Computation & Interpretation

-1	1
0	-2
2	0
1	-3
0	-1
-2	2
3	0
-1	1

Step 2

Take deviations from mean

Indian Statistical Institute

DESCRIPTIVE STATISTICS

Measures of Variation

Standard Deviation

Computation & Interpretation

1	1
0	4
4	0
1	9
0	1
4	4
3	0
1	1

Step 3

Square the deviations

Indian Statistical Institute

DESCRIPTIVE STATISTICS

Measures of Variation

Standard Deviation

1	1
0	4
4	0
1	9
0	1
4	4
3	0
1	1

Computation & Interpretation

Step 4

Sum the square of deviation

Sum of Squares = 40

Indian Statistical Institute

DESCRIPTIVE STATISTICS

Measures of Variation

Standard Deviation

1	1
0	4
4	0
1	9
0	1
4	4
3	0
1	1

Computation & Interpretation

Step 5

Variance = Average of the sum of square deviation

Variance = 40 / (16 – 1) = 2.667

Std Deviation = √Variance

= √2.667

= 1.633

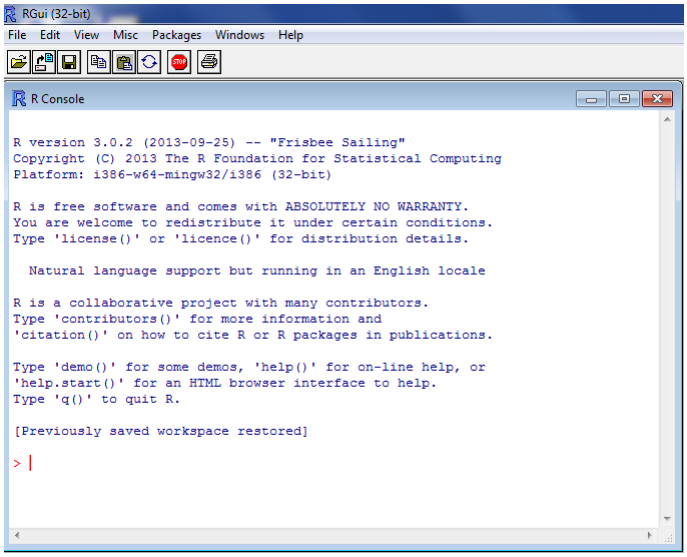
Introduction *to* R & R Studio

R INSTALLATION

1. Download R software from http://cran.r-project.org/bin/windows/base/
2. Run the R set up (exe) file and follow instructions

R INSTALLATION

3. Double click on the R icon in the desktop and R window will open

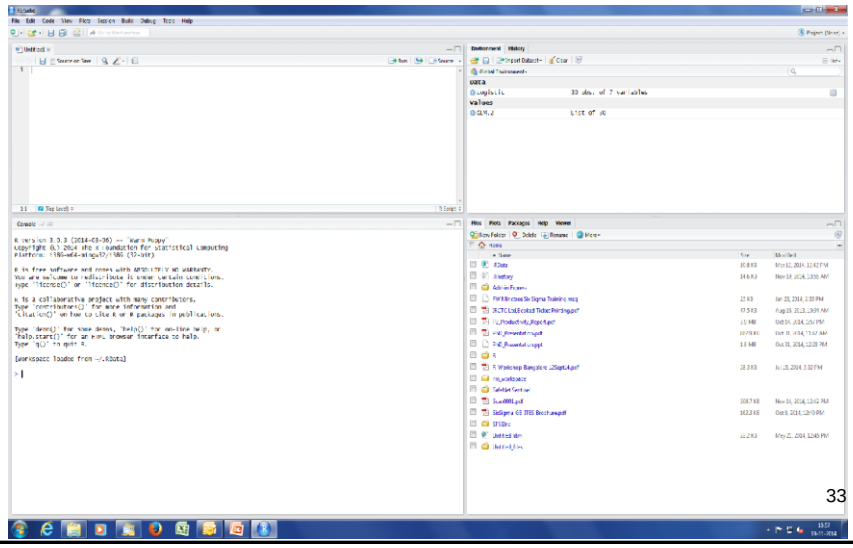


R INSTALLATION

4. Download R Studio from http://www.rstudio.com/
5. Run R studio set up file and follow instructions

R INSTALLATION

6. Click on R studio icon, R Studio IDE Studio will load



33

DESCRIPTIVE STATISTICS
***using* R**

34

DESCRIPTIVE STATISTICS

Exercise 1: The monthly credit card expenses of an individual in 1000 rupees is given in the file Monthly_Expenses.csv.

a. Read the dataset to R studio

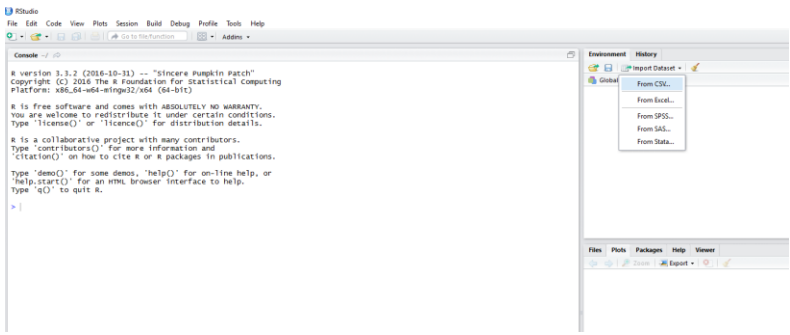
b. Compute mean, median minimum, maximum, range, variance, standard deviation, skewness, kurtosis and quantiles of Expenses

c. Compute default summary of Monthly Expenses

d. Draw Histogram of Monthly Expenses

DESCRIPTIVE STATISTICS

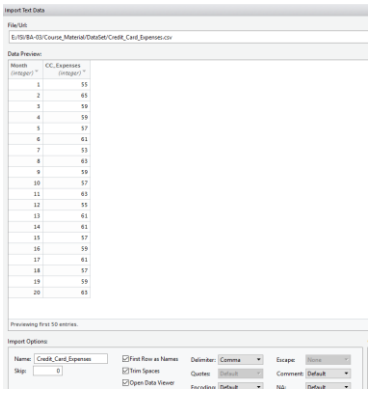
Reading a csv file to R Studio



The file open dialog box will pop up
Browse to the file

DESCRIPTIVE STATISTICS

Reading a csv file to R Studio



Click Import button
R studio will read the data set to a data frame with specified name

DESCRIPTIVE STATISTICS

Reading a csv file to R Studio : Source code

```
> Monthly_Expenses <- read.csv("D:/Infosys/DataSetsMonthly_Expenses.csv")
```

To change the name of the data set to : mydata

```
> mydata = Monthly_Expenses
```

To display the contents of the data set

```
> print(mydata)
```

To read a particular column or variable of data set to a ne variable

Example: Read Expenses to expenses

```
>expenses = mydata$Expenses
```

DESCRIPTIVE STATISTICS

Operators - Arithmetic

Operator	Description
+	addition
-	subtraction
*	multiplication
/	division
^ or **	exponentiation
x %% y	modulus (x mod y) 5%%2 is 1
x %/% y	integer division 5%/%2

39

DESCRIPTIVE STATISTICS

Operators - Logical

Operator	Description
<	less than
<=	less than or equal to
>	greater than
>=	greater than or equal to
==	exactly equal to
! =	not equal to
!x	Not x
x | y	x OR y
x & y	x AND y
isTRUE(x)	test if X is TRUE

40

DESCRIPTIVE STATISTICS

Descriptive Statistics

Computation of descriptive statistics for variable CC

Function	Code	Value
Mean	> mean(expenses)	59.2
Median	> median(expense)	59
Standard deviation	> sd(expense)	3.105174
Variance	> var(expenses)	9.642105
Minimum	> min(expenses)	53
Maximum	> max(expenses)	65
Range	> range(expenses)	53 65

DESCRIPTIVE STATISTICS

Descriptive Statistics

Function	Code
Quantile	> quantile(expenses)

Output					
Quantile	0%	25%	50%	75%	100%
Value	53	57	59	61	65

Function	Code
Summary	>summary(expenses)

Output					
Minimum	Q1	Median	Mean	Q3	Maximum
53	57	59	59.2	61	65

DESCRIPTIVE STATISTICS

Descriptive Statistics

Function	Code
describe	> libray(psych) > describe(expenses)

Output	
Statistics	Values
n	20
mean	59.2
sd	3.11
median	59
trimmed	59.25
mad	2.97
min	53
max	65
range	12
skew	-0.08
kurtosis	-0.85
se	0.69

43

DESCRIPTIVE STATISTICS

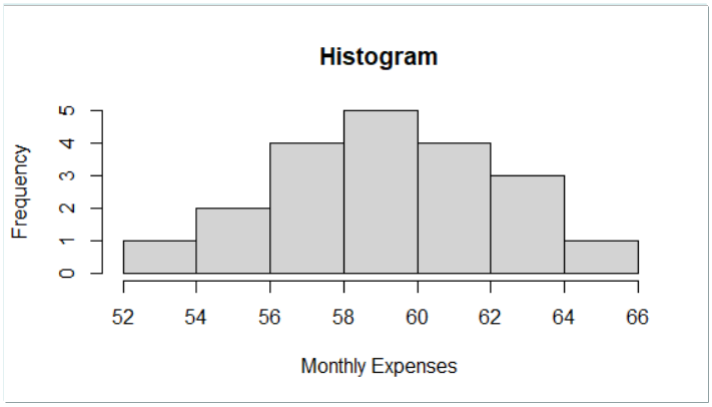
Graphs

Graph	Code
Histogram	> hist(expenses)
Histogram colour (“Blue”)	> hist(expense,col="blue")
Dot plot	> dotchart(expense)
Box plot	> boxplot(expense)
Box plot colour	> boxplot(expense, col="dark green")

44

DESCRIPTIVE STATISTICS

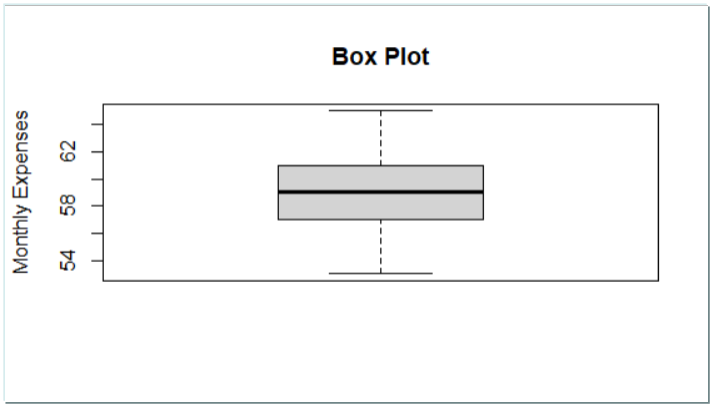
Histogram : Variable - expenses



45

DESCRIPTIVE STATISTICS

Box plot : Variable - expenses



46

R INTRODUCTION

About R

Case-sensitive, interpreted language.

Enter commands one at a time at the command prompt (>) or run a set of commands from a source file.

Data types: vectors (numerical, character, logical), matrices, data frames, and lists.

Most functionality is provided through built-in and user-created functions and all data objects are kept in memory during an interactive session.

Basic functions are available by default. Other functions are contained in packages or libraries that can be attached to a current session as needed.

47

R INTRODUCTION

Getting help

Getting general help

```
>help.start()
```

Help on a particular function say mean

```
> help("mean")
> ?mean
```

48

R INTRODUCTION

Packages

Packages are collections of R functions, data, and compiled code in a well-defined format.

The directory where packages are stored is called the library.

R comes with a standard set of packages.

Others are available for download and installation.

Once installed, they have to be loaded into the session to be used

To see all packages installed

```
> library()
```

To see all packages currently loaded

```
> search()
```

49

R INTRODUCTION

Data types

Scalars, vectors (numerical, character, logical), matrices, data frames, and lists.

Vector: Entering data using keyboard

```
>mydata = c(1.2, 2.3, 3.4, 4.5, 5.7)
```

Refer to elements of a vector (say 2nd, 5th element of mydata)

```
> mydata[c(2,5)]
```

Martix: All columns in a matrix must have same data type

```
> mydata = c(1:20)
> mymatrix = matrix(mydata, nrow = 5, ncol = 4, byrow = TRUE)
> mymatrix = matrix(mydata, nrow = 5, ncol = 4, byrow = FALSE)
```

50

R INTRODUCTION

Martix: Identifying column, row, elements

> mymatrix[,4]

3rd column of a matrix

> mymatrix[3,]

Rows 1 to 3 and columns 2 to 4 of a matrix

> mymatrix[1:3,2:4]

Data frame: Different columns can have different data types

> mydata = as.data.frame(mymatrix)

51

R INTRODUCTION

Data frame: Identifying column, row, elements

> mydata[,3]

> mydata[2,]

> mydata[1:3, 2:4]

> mydata["V1"]

> mydata$V2

52

R INTRODUCTION

Data Operations

```
> x1 = c(1:10)
> x2 = c(11:20)
```

Adding variables

```
> y = x1 + x2
```

Difference

```
> y = x1- x2
```

53

R INTRODUCTION

Data Operations

Multiplication

```
> y = x1*x2
```

Division

```
> y = x1/x2
```

Mathematical functions

```
> exp(x1)
> log(x1)
> log10(x1)
> sqrt(x1)
> abs(x1)
> factorial(x1)
```

54

Indian Statistical Institute

DESCRIPTIVE STATISTICS

Exercise 2: The data on productivity (number of tasks completed), developer experience (1: Experienced, 2: Fresher), Code reuse (1:High, 2: Low) and usage of knowledge repository usage (1: High , 2: Low) of an technical support process are given in file Productivty.csv.

a. Import the file to R Studio
b. Copy first 20 records from the file to another dataset and save it as a csv file
c. Compute descriptive summary of variable Productivity
d. Convert the variables developer experience, code reuse & knowledge repository usage to categorical (factor)
e. Check whether the average productivity varies with developer experience?
f. Check whether the average productivity vary with code reuse?
g. Check whether the average productivity vary with knowledge repository usage?
h. Compute the aggregate average of productivity with developer experience & code reuse?
i. Compute the aggregate average of usage with all three factors?

55

Indian Statistical Institute

DESCRIPTIVE STATISTICS

Exercise 2: The data on productivity (number of tasks completed), developer experience (1: Experienced, 2: Fresher), Code reuse (1:High, 2: Low) and usage of knowledge repository usage (1: High , 2: Low) of an technical support process are given in file Productivty.csv.

Reading dataset to variable: mydata

```
>mydata = Productivty
```

Copying first 20 rows to a new variable: mynewdata

```
> mynewdata = mydata[1:20,1:5]
```

Saving mynewdata to a csv file named mynewdata

```
> write.csv(mynewdata, "E:/ISI/BA-03/Course_Material/newdata.csv")
> write_excel_csv(mynewdata, "E:/ISI/BA-03/Course_Material/newdata.xls")
```

56

DESCRIPTIVE STATISTICS

Exercise 2: The data on productivity (number of tasks completed), developer experience (1: Experienced, 2: Fresher), Code reuse (1:High, 2: Low) and usage of knowledge repository usage (1: High , 2: Low) of an technical support process are given in file Productivty.csv.

Reading variable productivity to a new variable: prodn

> prodn = mydata$Productivity

Computing descriptive statistics for variable : Productivity

> summary(prodn)

Minimum	Q1	Median	Mean	Q3	Maximum
20	32.5	60	64.83	86.25	150

DESCRIPTIVE STATISTICS

Exercise 2: The data on productivity (number of tasks completed), developer experience (1: Experienced, 2: Fresher), Code reuse (1:High, 2: Low) and usage of knowledge repository usage (1: High , 2: Low) of an technical support process are given in file Productivty.csv.

Converting variables developer expereince, code reuse knowledge repository usage to factors

> experience = factor(mydata$Developer_Experience)

> reuse = factor(mydata$Code_Reuse)

> know_rep = factor(mydata$Knowledge_Repository_Usage)

Computing average productivity for different experience

> aggregate(prodn, by = list(experience), FUN = mean)

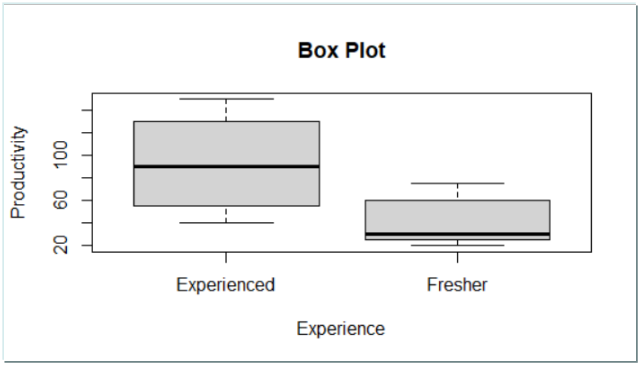
Group	Experience	Average Productivity
1	Experienced	88.67
2	Fresher	41.00

DESCRIPTIVE STATISTICS

Exercise 2: The data on productivity (number of tasks completed), developer experience (1: Experienced, 2: Fresher), Code reuse (1:High, 2: Low) and usage of knowledge repository usage (1: High , 2: Low) of an technical support process are given in file Productivty.csv.

.Box plot of Credit Card usage by sex

```
> boxplot(prodn ~ experience)
```



DESCRIPTIVE STATISTICS

Exercise 2: The data on productivity (number of tasks completed), developer experience (1: Experienced, 2: Fresher), Code reuse (1:High, 2: Low) and usage of knowledge repository usage (1: High , 2: Low) of an technical support process are given in file Productivty.csv.

Computing aggregate average of productivity for different experience and reuse

```
> aggregate(prodn ~ experience + reuse, FUN = mean)
```

Developer Experience	Code Reuse	Average Productivity
Experienced	High	96.36
Fresher	High	52.50
Experienced	Low	67.5
Fresher	Low	33.33

DESCRIPTIVE STATISTICS

Exercise 2: The data on productivity (number of tasks completed), developer experience (1: Experienced, 2: Fresher), Code reuse (1:High, 2: Low) and usage of knowledge repository usage (1: High , 2: Low) of an technical support process are given in file Productivty.csv.

Computing aggregate average of productivty by 3 factors

```
aggregate(prodn ~ experience + reuse + know_rep, FUN = mean)
```

Experience	Reuse	Know_Rep Usage	Average Productivity
Experienced	High	High	130
Fresher	High	High	75
Experienced	Low	High	80
Experienced	High	Low	56
Fresher	High	Low	48
Experienced	Low	Low	55
Fresher	Low	Low	33.33

61

Indian Statistical Institute

DESCRIPTIVE STATISTICS

Exercise 2: The data on productivity (number of tasks completed), developer experience (1: Experienced, 2: Fresher), Code reuse (1:High, 2: Low) and usage of knowledge repository usage (1: High , 2: Low) of an technical support process are given in file Productivty.csv.

Computing aggregate summary of productivity by 3 factors

```
> aggregate(prodn ~ experience + reuse + know_rep, FUN = summary)
```

Experience	Reuse	Usage	Productivity					
			Minimum	Q1	Median	Mean	Q3	Maximum
Experienced	High	High	90	130	135	130	140	150
Fresher	High	High	40	40	75	75	75	75
Experienced	Low	High	50	60	80	80	85	90
Experienced	High	Low	30	40	50	56	60	90
Fresher	High	Low	30	30	60	48	60	60
Experienced	Low	Low	80	82.5	55	55	57.5	60
Fresher	Low	Low	20	20	30	33.33	40	60

62

DESCRIPTIVE STATISTICS

Exercise 3: In IT service provider has conducted a customer satisfaction survey. The four important questions asked are given below: The respondents have to answer each question in a 7 point scale with 1: least satisfied and 7: most satisfied. The data is given in Csat_Freq_table.csv

Q1. Considering all aspects of your interactions, you are very satisfied with your experience with our company

Q2. You will definitely continue to use our company for your future needs

Q3. If a professional associate/colleague has a need for IT consulting and solutions / IT Infrastructure Services/ IT Engineering Services, you will definitely recommend our company

Q4. You believe that our company delivers the best value for money

a. Summarize each question responses using frequency table

b. Pictorially represent the responses to each question using pie chart and bar chart?

DESCRIPTIVE STATISTICS

Exercise 3: In IT service provider has conducted a customer satisfaction survey. The four important questions asked are given below: The respondents have to answer each question in a 7 point scale with 1: least satisfied and 7: most satisfied. The data is given in Csat_Freq_table.csv

Reading the data set to variable: mydata

```
> mydata = CSat_Freq_Table
```

Computing Frequency table for Q4

```
> mytable = table(mydata$q4)
```

```
> print(mytable)
```

Rating	Frequency
2	1
3	13
4	35
5	11
6	108
7	11

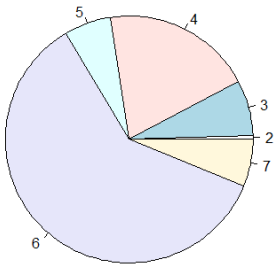
Indian Statistical Institute

DESCRIPTIVE STATISTICS

Exercise 3: In IT service provider has conducted a customer satisfaction survey. The four important questions asked are given below: The respondents have to answer each question in a 7 point scale with 1: least satisfied and 7: most satisfied. The data is given in Csat_Freq_table.csv

Creating pie chart for Q4

> pie(mytable)



65

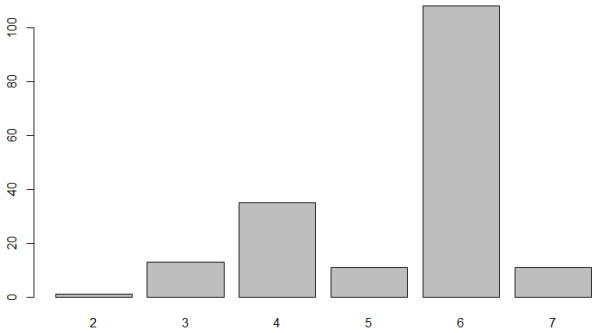
Indian Statistical Institute

DESCRIPTIVE STATISTICS

Exercise 3: In IT service provider has conducted a customer satisfaction survey. The four important questions asked are given below: The respondents have to answer each question in a 7 point scale with 1: least satisfied and 7: most satisfied. The data is given in Csat_Freq_table.csv

Creating bar chart for Q4

> barplot(mytable)



66

DATA PREPROCESSING

DATA PREPROCESSING

1. Missing value replenishment
2. Merging data files
3. Appending the data files
4. Transformation or normalization

Indian Statistical Institute

MISSING VALUE HANDLING

Example: The data on sprint productivity along with software development process variables are given in Preprocesing_Data1 file. Handle the missing values

69

Indian Statistical Institute

MISSING VALUE HANDLING

Example: Sprint productivity data

Read data and variables to R

```
> mydata = Preprocessing_Data_I
```

Two Options

1. Delete all records with missing values
2. Replace the missing values with a suitable statistics

70

Indian Statistical Institute

MISSING VALUE HANDLING

Option 1: Discard all records with missing values

```
>newdata = na.omit(mydata)
>write.csv(newdata,"E:/ISI_Mumbai/newdata.csv")
```

Disadvantage
Number of records will reduce
Some of the data already available will become unusable

71

Indian Statistical Institute

MISSING VALUE HANDLING

Option 2: Replace the missing values with variable mean, median, etc

```
# Computing the mssing vlaues of review time with mean
review_time = Review_Time
rev_time_mean = mean(Review_Time, na.rm = TRUE)
review_time[is.na(Review_Time)] = rev_time_mean
review_time

# Computing the mssing vlaues of test coverage with median
Test_coverage = Test_Coverage
test_cov_median = median(Test_Coverage, na.rm = T)
Test_coverage[is.na(Test_coverage)] = test_cov_median
Test_coverage
```

72

MISSING VALUE HANDLING

Option 2: Replace the missing values with variable mean, median, etc

```
# Computing the mssing vlaues of review coverage with 100
review_coverage = Review_Coverage
review_coverage[is.na(review_coverage)] = 100
review_coverage


# Preparing cleaned up data
newdata = cbind(Reviewer_Skill, Review_Type, Domain_Knowledge,
            review_time, Test_coverage, review_coverage,
            Reuse, Review_Rate, Sprint_Prod)
```


DATA MERGING

Exercise: The data is collected for optimizing a mailing campaign. The features are given in Mail_Repond_Features.csv file and the response is given in Mail_Respond_Response.txt file. Can you merge the two files into a single data set?

Read the files

```
> Features = Mail_Respond_Features
> Response = Mail_Respond_Response
```

Merge the files by “SL_No” field

```
>mydata = merge(Features, Response, by = “SL_No”)
```

DATA APPEND

Exercise: The data is collected from a software development process to study the relationship between the sprint productivity and process features. The data collected is given in two files namely Preprocessing_Data_I and Preprocessing_Data_II. Can you append the second file with the first one?

Read the files
>SP_I = Preprocessing_Data_I
> SP_II = Preprocessing_Data_II

Append class1 with class2
>mydata = rbind(SP_I, SP_II)

TRANSFORMATION / NORMALIZATION

z transform:

Transformed data = (Data – Mean) / SD

Exercise : The TAT data of a tech support process is given in TAT file. Normalize the variables in the TAT data?

```
Read the files
mydata = TAT

# Data summary
library(psych)
describe(mydata)

# z transform
myzdata = scale(mydata)
describe(myzdata)
```

TRANSFORMATION / NORMALIZATION

Min-Max transform:

Transformed data = (Data – Minimum) / (Maximum – Minimum)

Exercise : The TAT data of a tech support process is given in TAT file. Normalize the variables in the TAT data?

```
# Computing minimum and maximum of data
mins = apply(mydata, 2, min)
Mins

maxs = apply(mydata, 2, max)
Maxs

# Making min-max transformation
tr_data = scale(mydata, center = mins, scale = maxs - mins)
describe(tr_data)
```

77

DATA VISUALIZATION

78

DATA VISUALIZATION

Indian Statistical Institute

Methodology for exploring the relationship between fields

Generally used to explore relationship between response and explanatory variables in supervised learning

Explanatory variable	Response	Plot
Continuous	Continuous	Scatter plot
Continuous	Categorical	Boxplot, Density plot

79

DATA VISUALIZATION

Indian Statistical Institute

Example:

The data on temperature, time, viscosity and yield are given in Chemical_Yield file.

1. Replace the missing values using imputation?
2. Explore the relationship of temperature, time and viscosity to yield graphically

80

DATA VISUALIZATION

Example:

```
# Remove SL No column
mydata = mydata[, c(-1)]

# Changing mydata to dataframe format
mydata = as.data.frame(mydata)

#importing caret library
library(caret)
```

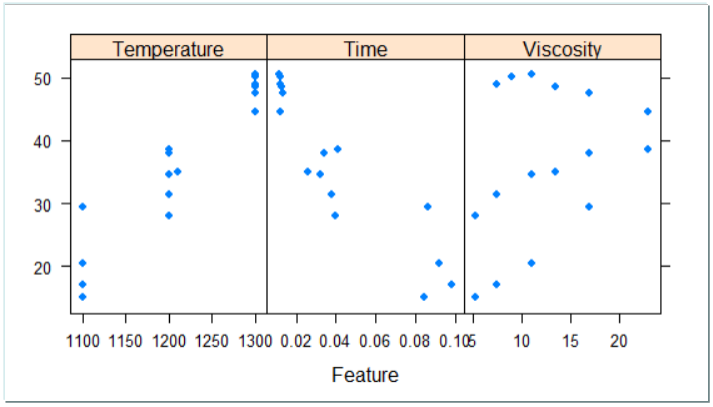

DATA VISUALIZATION

Example:

```
# seperating xand y varaibles
x = mydata[,1:3]
y = mydata[,4]

# Scatter plot matrix
featurePlot(x, y, plot = “scatter”, pch = 19)
```

DATA VISUALIZATION

Indian Statistical Institute



83

DATA VISUALIZATION

Indian Statistical Institute

Example:

The data on defect proneness of a tech support process is given in Defect_Proneness file. Explore the relationship between defect proneness with the process variables graphically

84

DATA VISUALIZATION

Example:

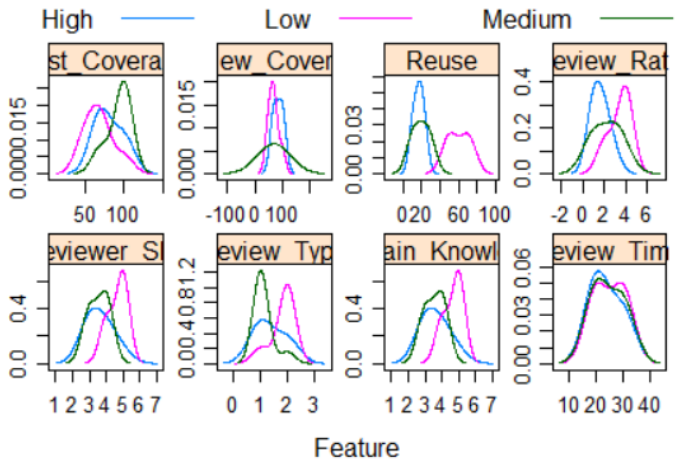
```
mydata = Defect_Proneness

# Changing mydata to dataframe format
mydata = as.data.frame(mydata)

# Importing caret library
library(caret)

# Seperating x and y varaibles
x = mydata[,1:8]
y = mydata[,9]
y = as.factor(y)
```


DATA VISUALIZATION

Example:

```
# Density plot
featurePlot(x, y, plot = "density", auto.key = list(columns = 3),
            scales = list(x = list(relation = "free"),
                          y = list(relation = "free")),
            adjust = 1.5, pch = "",
            layout = c(4,2))
```

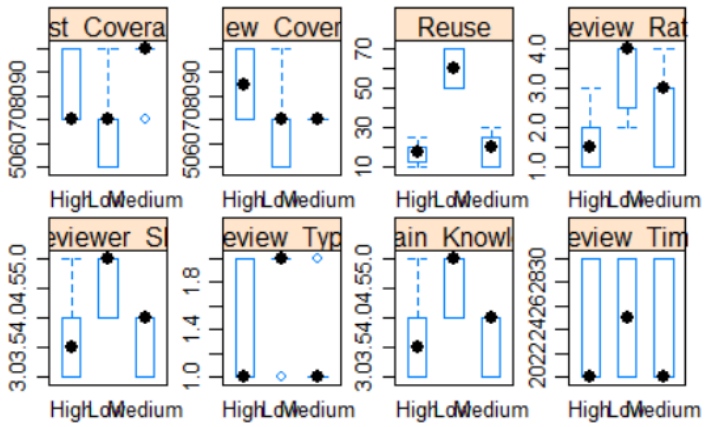
Density Plot:



Example:

```
# Box plot
featurePlot(x1, y, plot = “box", auto.key = list(columns = 3),
        scales = list(x = list(relation = "free"),
                y = list(relation = "free")),
        layout = c(4,2))
```

Box Plot:



TEST
of
HYPOTHESIS

TEST OF HYPOTHESIS

Introduction:

In many situations, it is required to accept or reject a statement or claim about some parameter

Example:

1. The average cycle time is less than 24 hours
2. The % rejection is only 1%

The statement is called the hypothesis

The procedure for decision making about the hypothesis is called hypothesis testing

Advantages

1. Handles uncertainty in decision making
2. Minimizes subjectivity in decision making
3. Helps to validate assumptions or verify conclusions

91

TEST OF HYPOTHESIS

Some of the commonly used hypothesis tests:

- Checking mean equal to a specified value ($mu = mu_0$)
- Two means are equal or not ($mu_1 = mu_2$)
- Two variances are equal or not ($sigma_1^2 = sigma_2^2$)

- Proportion equal to a specified value ($P = P_0$)
- Two Proportions are equal or not ($P_1 = P_2$)

92

TEST OF HYPOTHESIS

Null Hypothesis:

A statement about the status quo

One of no difference or no effect

Denoted by H0

Alternative Hypothesis:

One in which some difference or effect is expected

Denoted by H1

93

TEST OF HYPOTHESIS

Types of errors in hypothesis testing

The decision procedure may lead to either of the two wrong conclusions

Type I Error

Rejecting the null hypothesis H0 when it is true

Type II Error

Failing to reject the null hypothesis H0 when it is false

Alpha (Significance level) = Probability of making type I error

Beta = Probability of making type II error

Power = 1 – Beta : Probability of correctly rejecting a false null hypothesis

94

TEST OF HYPOTHESIS

Hypothesis Testing: General Procedure

1. Formulate the null hypothesis H0 and the alternative hypothesis H1
2. Gather evidence (data collection)
3. Based on evidence take a decision to accept or reject H0

TEST OF HYPOTHESIS

Methodology demo: To Test Mean = Specified Value (mu = mu_0)

Suppose we want to test whether mean of a process characteristic is 5 based on the following sample data from the process

4	4	5	5	6
5	4.5	6.5	6	5.5

Calculate the mean of the sample, xbar = 5.15

Compare xbar with specified value 5

or xbar - specified value = xbar - 5 with 0

If xbar - 5 is close to 0

then conclude mean = 5

else mean $\neq$ 5

TEST OF HYPOTHESIS

Methodology demo : To Test Mean = Specified Value (mu = mu_0)

Consider another set of sample data. Check whether mean of the process characteristic is 500

400	400	500	500	600
500	450	650	600	550

Mean of the sample, xbar = 515

xbar - 500 = 515 - 500 = 15

Can we conclude mean ≠ 500?

Conclusion:

Difficult to say mean = specified value by looking at xbar - specified value alone

TEST OF HYPOTHESIS

Methodology demo: To Test Mean = Specified Value (mu = mu_0)

Test statistic is calculated by dividing (xbar - specified value) by a function of standard deviation

To test Mean = Specified value

Test Statistic t_0 = (xbar - Specified value) / (SD / √n)

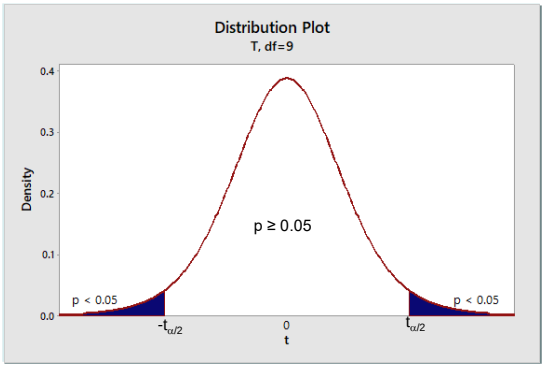
$$t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

If test statistic is close to 0, conclude that Mean = Specified value

To check whether test statistic is close to 0, find out p value from the sampling distribution of test statistic

TEST OF HYPOTHESIS

Methodology demo: To Test Mean = Specified Value (mu = mu_0)



If test statistic t_0 is close to 0 then p will be high ($p \geq 0.05$)

If test statistic t_0 is not close to 0 then p will be small ($p < 0.05$)

If p is high , $p \geq 0.05$ (with alpha = 0.05), conclude that $t \approx 0$, then

Mean = Specified Value, H0 is not rejected

99

TEST OF HYPOTHESIS

To Test Mean = Specified Value (mu = mu_0)

Example: Suppose we want to test whether mean of the process characteristic is 5 based on the following sample data

4	4	5	5	6
5	4.5	6.5	6	5.5

H0: Mean = 5

H1: Mean $\neq$ 5

Calculate xbar = 5.15

SD = 0.8515

n = 10

Test statistic t_0 = (xbar - 5)/(SD / $\sqrt{n}$) = (5.15 - 5) / (0.8515 / $\sqrt{10}$) = 0.5571

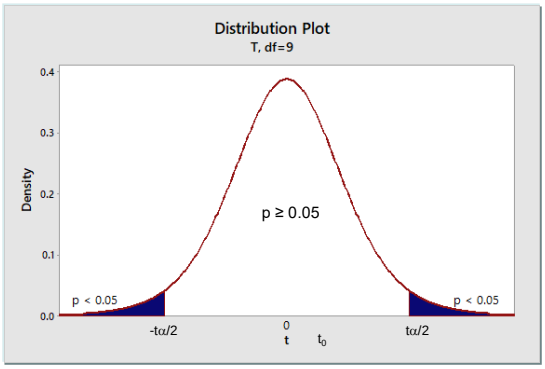
Critical value = ± 2.263 corresponding to sample size of 10

100

TEST OF HYPOTHESIS

Example: To Test Mean = Specified Value (mu = mu_0)

$t_0 = 0.5571$



p = 0.59 ≥ 0.05, hence Mean = Specified value = 5.

H0: Mean = 5 is not rejected

TEST OF HYPOTHESIS

Hypothesis Testing: Steps

1. Formulate the null hypothesis H0 and the alternative hypothesis H1
2. Select an appropriate statistical test and the corresponding test statistic
3. Choose level of significance alpha (generally taken as 0.05)
4. Collect data and calculate the value of test statistic
5. Determine the probability associated with the test statistic under the null hypothesis using sampling distribution of the test statistic
6. Compare the probability associated with the test statistic with level of significance specified

TEST OF HYPOTHESIS

One sample t test

Exercise 1 : A company claims that on an average it takes only 40 hours or less to process any purchase order. Based on the data given below, can you validate the claim? The data is given in PO_Processing.csv

TEST OF HYPOTHESIS

One sample t test

Exercise 1 : A company claims that on an average it takes only 40 hours or less to process any purchase order. Based on the data given below, can you validate the claim? The data is given in PO_Processing.csv

Reading data to mydata

```
> mydata = PO_Processing$Processing_Time
```

Performing one sample t test

```
> t.test(mydata, alternative = 'greater', mu = 40)
```

Statistics	Value
t	3.7031
df	99
P value	0.0001753
Critical value	1.984

TEST OF HYPOTHESIS

One sample t test

Exercise 2 : A computer manufacturing company claims that on an average it will respond to any complaint logged by the customer from anywhere in the world within 24 hours. Based on the data, validate the claim? The data is given in Compaint_Response_Time.csv

Response Time	
24	26
31	27
29	24
26	23
28	27
26	28
29	27
29	23
27	27
31	23
25	25
29	27
29	26
25	28
26	27

TEST OF HYPOTHESIS

To Test Two Means are Equal:

Null hypothesis H0: $Mean_1 = Mean_2$ ($mu_1 = mu_2$)

Alternative hypothesis H1: $Mean_1 \neq Mean_2$ ($mu_1 \neq mu_2$)

or

H1: $Mean_1 > Mean_2$ ($mu_1 > mu_2$)

or

H1: $Mean_1 < Mean_2$ ($mu_1 < mu_2$)

TEST OF HYPOTHESIS

To Test Two Means are Equal: Methodology

Calculate both sample means xbar1 & xbar2

Calculate SD1 & SD2

Compare xbar1 with xbar2

Or xbar1 - xbar2 with 0

Calculate test statistic t_0 by dividing (xbar1 – xbar2) by a function of SD1 & SD2

t_0 = (xbar1 – xbar2) / (Sp $\sqrt{}$ ((1/n1)+(1/n2)))

Calculate p value from t distribution

If p ≥ 0.05 then H0: $Mean_1 = Mean_2$ is not rejected

TEST OF HYPOTHESIS

To Test Two Means are Equal: Methodology

When variances are equal

$$t_0 = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

When variances are not equal

$$t_0 = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

TEST OF HYPOTHESIS

To Test Two Variances are Equal: Methodology ($Sigma_1^2 = Sigma_2^2$)

Null hypothesis

H0: $Sigma_1^2 = Sigma_2^2$

Alternative hypothesis

H1: $Sigma1^2 \neq Sigma_2^2$

Calculate standard deviations of both the samples S1 & S2

Calculate test statistic $F = S1^2 / S2^2$

If F is close to 1, then $S1^2$ more or less equal to $S2^2$

Calculate p from F distribution.

If $p \geq 0.05$ (with alpha = 0.05), then

H0: $Sigma_1^2 = Sigma_2^2$ is not rejected

TEST OF HYPOTHESIS

Two sample t test

Exercise 1: A super market chain has introduced a promotional activity in its selected outlets in the city to increase the sales volume. Based on the data given below, check whether the promotional activity resulted in increasing the sales. The outlets where promotional activity introduced are denoted by 1 and others by 2? The data is given in Sales_Promotion.csv

Outlet	Sales	Outlet	Sales
1	1217	2	1731
1	1416	2	1420
1	1381	2	1065
1	1413	2	1612
1	1800	2	1361
1	1724	2	1259
1	1310	2	1470
1	1616	2	622
1	1941	2	1711
1	1792	2	2315
1	1453	2	1180
1	1780	2	1515

TEST OF HYPOTHESIS

Two sample t test

Exercise 1: A super market chain has introduced a promotional activity in its selected outlets in the city to increase the sales volume. Based on the data given below, check whether the promotional activity resulted in increasing the sales. The outlets where promotional activity introduced are denoted by 1 and others by 2?

Reading data to mydata

```
> mydata = Sales_Promotion
> Outlet = mydata$Outlet
> Sales = mydata$Sales
```

Checking variances are equal or not

```
> var.test(sales ~ outlet)
```

Statistic	Value
F	0.31959
p_value	0.0713

111

TEST OF HYPOTHESIS

Two sample t test

Exercise 1: A super market chain has introduced a promotional activity in its selected outlets in the city to increase the sales volume. Based on the data given below, check whether the promotional activity resulted in increasing the sales. The outlets where promotional activity introduced are denoted by 1 and others by 2?

2 sample t Test

```
t.test(sales ~ outlet, alternative = "greater", var.equal = TRUE)
```

Statistics	Value
t	0.9625
df	22
P value	0.1731

112

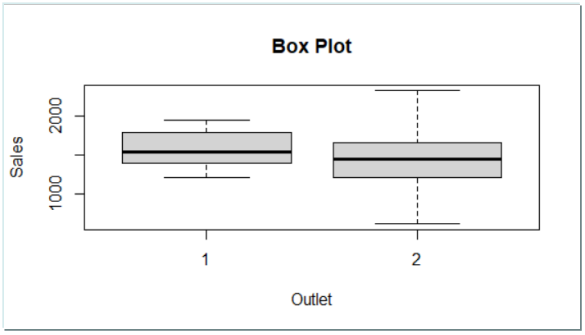
TEST OF HYPOTHESIS

Two sample t test

Exercise 1: A super market chain has introduced a promotional activity in its selected outlets in the city to increase the sales volume. Based on the data given below, check whether the promotional activity resulted in increasing the sales. The outlets where promotional activity introduced are denoted by 1 and others by 2?

Box Plot

```
> boxplot(Sales~Outlet)
```



113

TEST OF HYPOTHESIS

Two sample t test

Exercise 2: A bpo company have developed a new method for better utilization of its resources. 10 observations on utilization from both methods are given below: Check whether the mean utilization for both methods are same or not? Data is given in Utilization.csv.

Method	Utilization	Method	Utilization
Old	89.5	New	89.5
Old	90	New	91.5
Old	91	New	91
Old	91.5	New	89
Old	92.5	New	91.5
Old	91	New	92
Old	89	New	92
Old	89.5	New	90.5
Old	91	New	90
Old	92	New	91

114

TEST OF HYPOTHESIS

Paired t test:

A special case of two sample t test

When observations on two groups are collected in pairs

Each pair of observation is taken under homogeneous conditions

Procedure

Compute d: difference in paired observations

Let difference in means be $\mu_D = \mu_1 - \mu_2$

Null hypothesis H0: $\mu_D = 0$

Alternative hypothesis H1: $\mu_D \neq 0$ or $\mu_D > 0$ or $\mu_D < 0$

Test statistics t0 = $\dfrac{\bar{d} - 0}{s_d / \sqrt{n}}$

Reject H0 if p – value < 0.05

115

TEST OF HYPOTHESIS

Paired t test: Exercise 1

The manager of a fleet of automobiles is testing two brands of radial tires. He assigns one tire of each brand at random to the two rear wheels of eight cars and runs the cars until the tire wear out. Is both brands have equal mean life? The data in kilometers is given in tires.csv

Brand 1	Brand 2
36925	34318
45300	42280
36240	35500
32100	31950
37210	38015
48360	47800
38200	37810
33500	33215

116

TEST OF HYPOTHESIS

Paired t test: Exercise 1

The manager of a fleet of automobiles is testing two brands of radial tires. He assigns one tire of each brand at random to the two rear wheels of eight cars and runs the cars until the tire wear out. Is both brands have equal mean life? The data in kilometers is given in tires.csv

Reading the file and variables

```
> mydata = Tires
> One = mydata$Brand.1
> Two = mydata$Brand.2
```

Paired t test

```
> t.test(One,Two, paired = TRUE)
```

Box Plot

```
> boxplot(mydata)
```

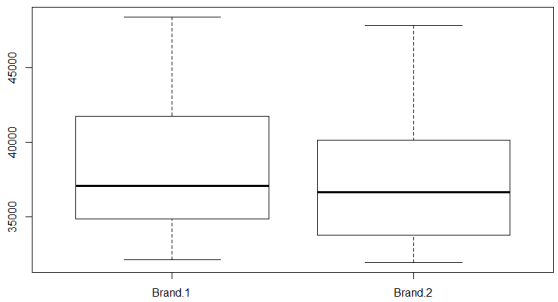
Statistics	Value
t	1.9039
df	7
P value	0.09863

TEST OF HYPOTHESIS

Paired t test: Exercise 1

The manager of a fleet of automobiles is testing two brands of radial tires. He assigns one tire of each brand at random to the two rear wheels of eight cars and runs the cars until the tire wear out. Is both brands have equal mean life? The data in kilometers is given in tires.csv

Box Plot



TEST OF HYPOTHESIS

Paired t test: Exercise 2

Ten individuals have participated in a diet – modification program to stimulate weight loss. Their weights (in kg) both before and after participation in the program is given in Diet.csv. On an average is the program successful? On an average whether the weight is reduced by 5 kg?

Subject	Before	After
1	88	85
2	97	88
3	112	100
4	91	86
5	85	79
6	95	89
7	98	90
8	112	100
9	133	126
10	141	129

TEST OF HYPOTHESIS

Paired t test: Exercise 2

Ten individuals have participated in a diet – modification program to stimulate weight loss. Their weights (in kg) both before and after participation in the program is given in Diet.csv. On an average is the program successful? On an average whether the weight is reduced by 5 kg?

R Code

```
before = mydata$Before
after = mydata$After

t.test(before, after, alternative = "greater", paired = T)
t.test(before, after, alternative = "greater", mu = 5, paired = T)
```

TEST OF HYPOTHESIS

Discrete Data: To Test Proportion is equal to Specified Value ($p = p_0$)

Null hypothesis H0: p = Specified Value ($p = p_0$)

Alternative hypothesis H1: $p \neq$ Specified Value ($p \neq p_0$)

or

H1: $p >$ Specified Value ($p > p_0$)

or

H1: $p <$ Specified Value ($p < p_0$)

TEST OF HYPOTHESIS

To Test Proportion is equal to a Specified Value: Methodology

Calculate sample proportion $\hat{p}$

Compare $\hat{p} = \text{specifiedvalue}(p_0)$

Or $\hat{p} - p_0 = 0$

Calculate test statistic z by dividing $\hat{p} - \text{specifiedvalue}$ by SD

$z0 = (\hat{p} - p_0) / \sqrt{(p_0 (1 - p_0)/n)}$

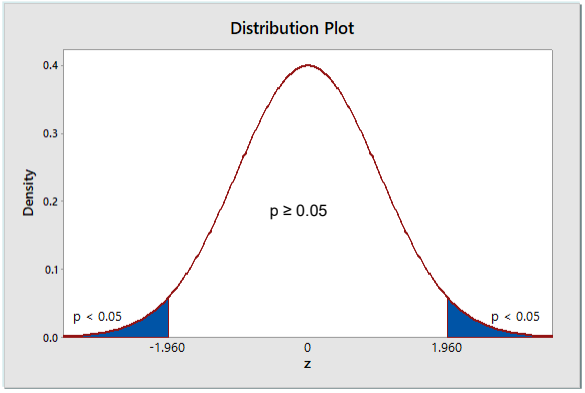
Calculate p value from z distribution

If p value ≥ 0.05 then H0: p = Specified Value is not rejected

TEST OF HYPOTHESIS

To Test Proportion is equal to a Specified Value: Methodology

If p value ≥ 0.05 then H0: p = Specified Value is not rejected



TEST OF HYPOTHESIS

One sample Proportion test

Exercise 1

A city branch of a bank claims that they are at least 99 % accurate on loan processing and at most only 1 % of loans are reworked. Validate the claim based on the data given in loan_processing.csv?

TEST OF HYPOTHESIS

One sample Proportion test

Exercise 1

A city branch of a bank claims that they are at least 99 % accurate on loan processing and at most only 1 % of loans are reworked. Validate the claim based on the data given in loan_processing.csv?

Reading the data and variables

```
> mydata = Loan_processing
```

Summarizing the data

```
> mytable = table(mydata)
> print(mytable)
```

Category	Count
Good	1482
Rework	31

TEST OF HYPOTHESIS

One sample Proportion test

Exercise 1

A city branch of a bank claims that they are at least 99 % accurate on loan processing and at most only 1 % of loans are reworked. Validate the claim based on the data given in loan_processing.csv?

One sample proportion test

```
> prop.test(mytable,alternative = 'less', p = 0.99)
```

Statistics	Value
X - squared	15.7715
df	1
p value	0.000

TEST OF HYPOTHESIS

One sample Proportion test

Exercise 2

A supply chain company claims that they deliver at least 98% of shipments without any damage. Based on the data in shipment.csv, validate the claim?

TEST OF HYPOTHESIS

To Test Two Proportion are equal : Methodology

Null Hypothesis H0: $p_1 = p_2$

Alternative Hypothesis H1: $p_1 \neq p_2$

or

H1: $p_1 > p_2$

or

H1: $p_1 < p_2$

TEST OF HYPOTHESIS

To Test Two Proportion are equal : Methodology

Calculate sample proportions $\hat{p}_1$ and $\hat{p}_2$

Check $\hat{p}_1 = \hat{p}_2$

Or $\hat{p}_1 - \hat{p}_2 = 0$

Calculate test statistic z_0 by dividing $\hat{p}_1 - \hat{p}_2$ by SD

$z_0 = (\hat{p}_1 - \hat{p}_2) / \sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n_2)}$

Calculate p value from z distribution

If p value ≥ 0.05 then H0: $p_1 = p_2$ is not rejected

TEST OF HYPOTHESIS

Two Proportion Test: Exercise 1

A multinational company suspects that the orders processed in their Bangalore bpo center is better than that done at their Pune office. Validate the claim based on the order processing data?

TEST OF HYPOTHESIS

Two Proportion Test: Exercise 1

A multinational company suspects that the orders processed in their Bangalore bpo center is better than that done at their Manila office. Validate the claim based on the order processing data?

Reading the data and variables

```
> mydata = Order_Processing
```

Summarizing the data

```
> mytable = table(mydata)
> print(mytable)
```

Location	Defective	Good
Bangalore	6	551
Pune	14	430

TEST OF HYPOTHESIS

Two Proportion Test: Exercise 1

A multinational company suspects that the orders processed in their Bangalore bpo center is better than that done at their Manila office. Validate the claim based on the order processing data?

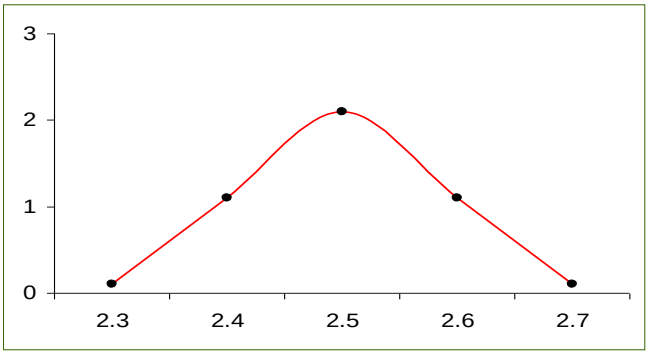
Two proportion test

```
> prop.test(mytable, alternative = 'less')
```

Statistics	Value
X - squared	4.4291
df	1
p value	0.01767

NORMALITY TEST

NORMAL DISTRIBUTION



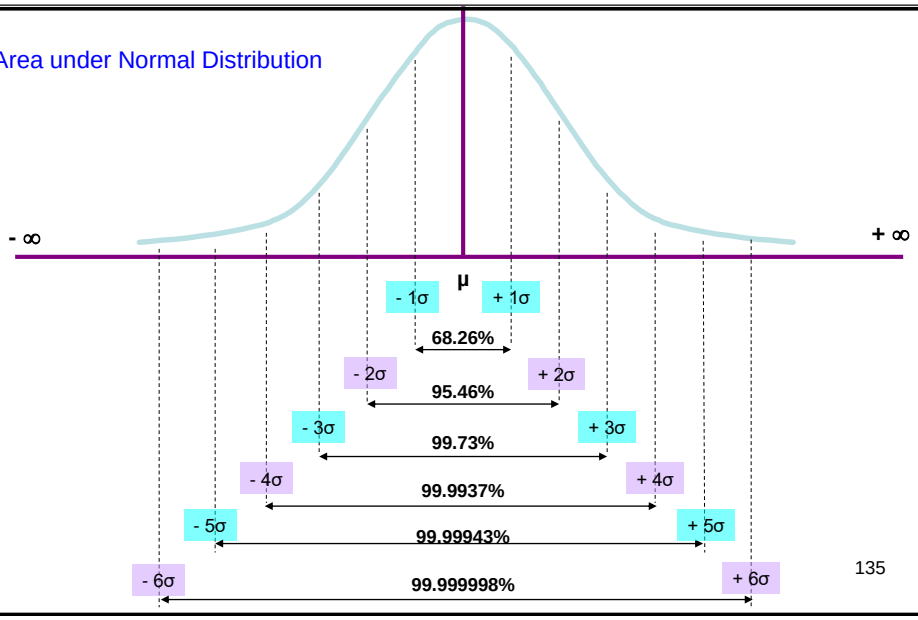
- Bell Shaped
- Symmetric
- Total Area under the curve is 1

Then : Normal Curve & Data follows Normal Distribution

Indian Statistical Institute

NORMAL DISTRIBUTION

Area under Normal Distribution

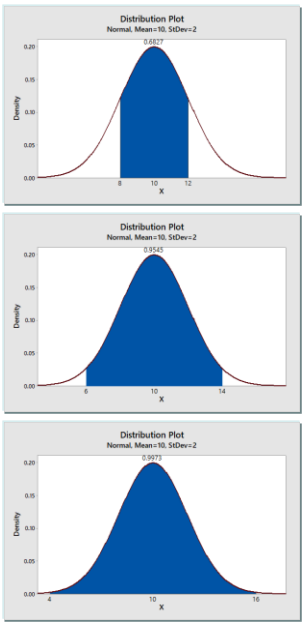


135

Indian Statistical Institute

NORMAL DISTRIBUTION

Distribution	Normal
Count	1000
Mean	10
Std Deviation	2
Mean ± 1 SD	8 to 12
Mean ± 2 SD	6 to 14
Mean ± 3 SD	4 to 16



136

NORMALITY TEST

Normality test

A methodology to check whether the characteristic under study is normally distributed or not

Two Methods

1. Quantile – Quantile (Q- Q) plot
2. Shapiro – Wilk test

NORMALITY TEST

Normality test - Quantile – Quantile (Q- Q) plot

- Plots the ranked samples from the given distribution against a similar number of ranked quantiles taken from a normal distribution
- If the sample is normally distributed then the line will be straight in the plot

NORMALITY TEST

Normality test – Shapiro – Wilk test

H0: Deviation from bell shape (normality) = 0

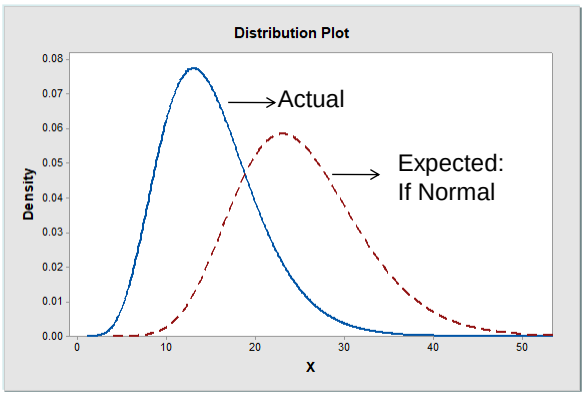
H1 : Deviation from bell shape ≠ 0

If p value ≥ 0.05 (5%), then H0 is not rejected, distribution is normal

139

NORMALITY TEST

Normality test – Shapiro – Wilk test



140

NORMALITY TEST

Normality test

Exercise 1 : The processing times of purchase orders is given in PO_Processing.csv. Is the processing time normally distributed?

NORMALITY TEST

Normality test

Exercise 1 : The processing times of purchase orders is given in PO_Processing.csv. Is processing time normally distributed?

Reading the data and variable

```
> mydata = PO_Processing
> PT = mydata$Processing_Time
```

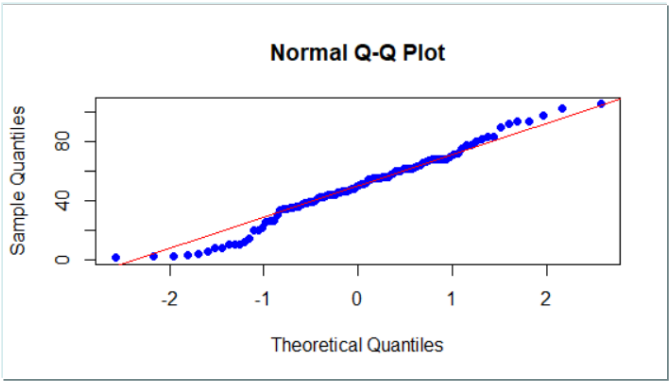
NORMALITY TEST

Normality test

Exercise 1 : The processing times of purchase orders is given in PO_Processing.csv. Is processing time normally distributed?

Normality Check using Normal Q – Q plot

```
qqnorm(PT, col = "blue", pch = 19)
qqline(PT, col = "red")
```



143

NORMALITY TEST

Normality test

Exercise 1 : The processing times of purchase orders is given in PO_Processing.csv. Is processing time normally distributed?

Normality Check using Shapiro – Wilk test

```
> shapiro.test(PT)
```

Statistics	Value
W	0.9804
p value	0.1418

144

NORMALITY TEST

Normality test

Exercise 2 : The time taken to respond to customer complaints is is given in Compaint_Response_Time.csv. Check whether the complaint response time follows normal distribution?

Response Time	
24	26
31	27
29	24
26	23
28	27
26	28
29	27
29	23
27	27
31	23
25	25
29	27
29	26
25	28
26	27

145

NORMALITY TEST

Normality test

Exercise 3 : The impurity level (in ppm) is routinely measured in an intermediate chemical process. The data is given in Impurity.csv. Check whether the impurity follows normal distribution?

146

ANALYSIS *of* VARIANCE

ANALYSIS OF VARIANCE

ANOVA

Analysis of Variance is a test of means for two or more populations

Partitions the total variability in the variable under study to different components

H0 = $Mean_1 = Mean_2 = - - - = Mean_k$

Reject H0 if p – value < 0.05

Example:

To study technology on coding productivity

ANALYSIS OF VARIANCE

One Way Anova : Example

An IT company wants to study whether the Technology have any effect of on the Coding Ratio. The data is collected given in Anova_Coding_Ratio file

	Technology		
	J2EE	Informix 4GL	C++
Coding ratio	1.34 1.89 1.35 2.07 2.41 3.06	3.20 2.81 4.52 4.40 4.75 5.19 3.42 6.80	2.30 1.91 1.40 1.48

149

ANALYSIS OF VARIANCE

One Way Anova : Example

Factor: Technology (A)

Levels : J2EE, Informix 4GL, C++

Response: Coding ratio

150

ANALYSIS OF VARIANCE

One Way Anova : Example

Step 1: Calculate the Sum, Average and Number of Responses for each level of the factor (Technology).

Level 1 Sum(A1):

Sum of all responses when Technology is at level 1 (J2EE)

= 1.34 + 1.89 + 1.35 + 2.07 + 2.41 + 3.06

= 12.12

nA1: Number of Responses with Technology is at level 1 (J2EE)

= 6

ANALYSIS OF VARIANCE

One Way Anova : Example

Step 1: Calculate the Sum, Average and Number of Responses for each level of the factor (Technology).

Level 1 Average:

Sum of all responses when Technology is at level 1 / Number of Responses with Technology is at level 1

= A1 / nA1 = 12.12 / 6 = 2.02

ANALYSIS OF VARIANCE

One Way Anova : Example

Step 1: Calculate the Sum, Average and Number of Responses for each level of the factor (Technology).

	Level 1 (J2EE)	Level 2 (Informix 4GL)	Level 3 (C ++)
Sum	A1: 12.12	A2: 35.09	A3: 7.09
Number	nA1: 6	nA2: 8	nA3: 4
Average	2.02	4.3863	1.7725

153

ANALYSIS OF VARIANCE

One Way Anova : Example

Step 2: Calculate the Grand Total (T)

T = Sum of all the Responses

= 1.34 + 1.89 + - - - + 1.40 + 1.48 = 547.3

Step 3: Calculate the Total Number of Responses (N)

N = 18

Step 4: Calculate the Correction Factor (CF)

CF = (Grand Total)^2 / Number of Responses

= T^2 / N = $(57.3)^2$ / 18 = 163.805

154

ANALYSIS OF VARIANCE

One Way Anova : Example

Step 5: Calculate the Total Sum of Squares (TSS)

TSS = Sum of Square of all the Responses - CF

$= 1.34^2 + 1.89^2 + - - - + 1.40^2 + 1.48^2 – 163.805$

$= 41.2918$

ANALYSIS OF VARIANCE

One Way Anova : Example

Step 6: Calculate the Sum of Square of Factor technology (A)

$SS_A = A1^2 / nA1 + A2^2 / nA2 + A3^2 / nA3 - CF$

$= 12.12^2 / 6 + 35.09^2 / 8 + 7.09^2 / 4 – 163.805$

$= 27.1579$

Step 7: Calculate the Sum of Square of Error (SS_e)

SS_e = Total Sum of Square - Sum of Square of Factors

$= TSS - SS_A = 41.2918 – 27.1579 = 14.1339$

ANALYSIS OF VARIANCE

One Way Anova : Example

Step 8: Calculate Degrees of Freedom (Df)

Total Df = Total Number of Responses - 1

= 18 - 1 = 17

Df of Factor technology

= Number of levels of the Factor - 1

= 3 - 1 = 2

Df of Error = Total Df - Df of Factors

= 17 - 2 = 15

ANALYSIS OF VARIANCE

One Way Anova : Example

Anova Table:

Source	SS	df	MS	F	p_value	F Crit
Between	27.1579	2	13.5790	14.4111	0.00032	3.68232
Within	14.1339	15	0.94226			
Total	41.2918	17	2.42893			

MS = SS / Df

F of a Factor = MS of Between / MS of Within

F table =finv (probability, df of factor, df of error) , probability = 0.05

P value = fdist (F, df of Factor, df of error)

ANALYSIS OF VARIANCE

One Way Anova : R Code

Reading data and variables to R

```
> mydata = Anova_Coding_Ratio
> technology = mydata$Technology
> coding = mydata$Coding_Ratio
```

Computing ANOVA table

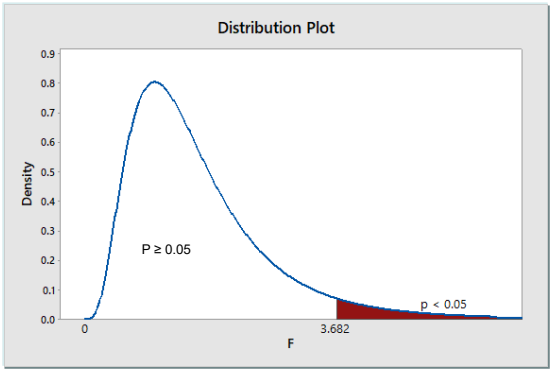
```
> mymodel = aov(coding ~ technology)
> summary(mymodel)
```


ANALYSIS OF VARIANCE

One Way Anova : Decision Rule

If p value < 0.05, then

The factor has significant effect on the process output or response.



Meaning:

When the factor is changed from 1 level to another level, there will be significant change in the response.

ANALYSIS OF VARIANCE

One Way Anova : Example Result

For factor Technology, p = 0.000 < 0.05

Conclusion:

Technology has significant effect on Coding Ratio.

Meaning:

The coding ratio is not same for different technologies like J2EE, Informix 4GL & C++

ANALYSIS OF VARIANCE

One Way Anova : Example Result

The expected coding ratio for different technologies under study is equal to level averages.

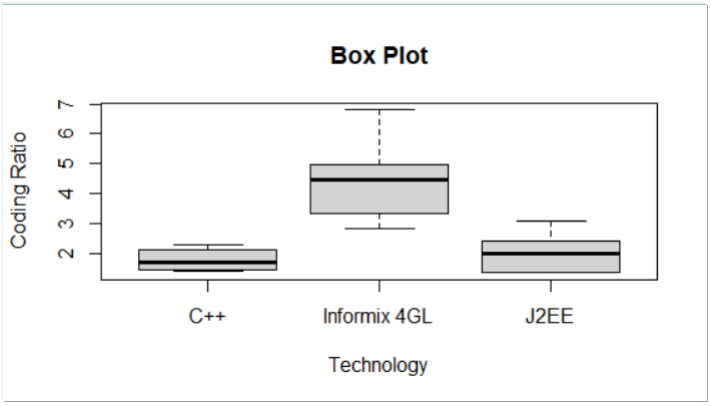
> aggregate(coding ~ technology, FUN = mean)

Technology	Expected coding ratio
J2EE	2.0200
Informix 4GL	4.3863
C ++	1.7725

ANALYSIS OF VARIANCE

One Way Anova : Example Result

```
> boxplot(coding ~ technology)
```

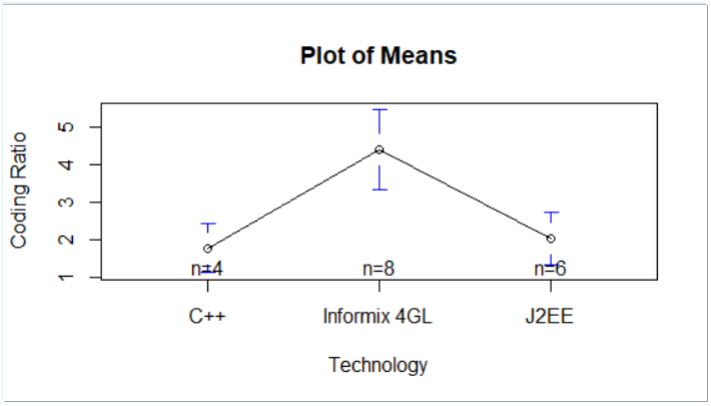


163

ANALYSIS OF VARIANCE

One Way Anova : Example Result

```
> library(gplots)
> plotmeans(coding ~ technology)
```



164

ANALYSIS OF VARIANCE

One Way Anova : Tukey's HSD Test

Used to do pair wise comparison between the levels of factors

R code

```
>TukeyHSD(mymodel)
```

Comparison	Mean difference	Lower	Upper	p value
Informix 4GL -C++	2.61375	1.069737	4.157763	0.001419
J2EE - C++	0.2475	-1.38003	1.875033	0.918053
J2EE - Informix 4GL	-2.36625	-3.72794	-1.00456	0.001128

165

ANALYSIS OF VARIANCE

Anova logic:

Two Types of Variations:

1. Variation within the level of a factor
2. Variation between the levels of factor

166

ANALYSIS OF VARIANCE

Theory Behind Anova:

Variation between the level of a Factor:

The effect of Factor.

Variation within the levels of a Factor:

The inherent variation in the process or Process Error.

	Technology		
	J2EE	Informix 4GL	C++
Coding ratio	1.34 1.89 1.35 2.07 2.41 3.06	3.20 2.81 4.52 4.40 4.75 5.19 3.42 6.80	2.30 1.91 1.40 1.48

167

ANALYSIS OF VARIANCE

Theory Behind Anova:

If the Variation between the levels of a Factor is significantly higher than the inherent variation

then the factor has significant effect on Response

To check whether a Factor is significant:

Compare Variation between levels with Variation within levels

168

ANALYSIS OF VARIANCE

Theory Behind Anova:

Measure of Variation between Levels: MS of the Factor

Measure of Variation within levels: MS Error

To check whether a Factor is significant:

Compare MS of Factor with MS Error

i.e. Calculate F = MS factor / MS Error

If F is very high, then the factor is significant.

ANALYSIS OF VARIANCE

Variation Within levels:

Ideally variation within all the levels should be same

To check whether variation within the levels are same or not

Do Bartlett's test

If p value ≥ 0.05, then variation within the levels are equal, otherwise not

R Code for Bartlett's test

```
> bartlett.test(coding ~ technology)
```

ANALYSIS OF VARIANCE

Variation Within levels:

Bartlett's Test result for sales revenue (location of TV sets) example

Bartlett's K^2 Statistic	df	p- value
4.6298	2	0.0988

Since p value = 0.0988 > 0.05, the variance within the levels are equal

ANALYSIS OF VARIANCE

Exercise 1: An insurance company wants to check whether the waiting time of customer at their single window operation across 4 cities is same or not. The data is given in Insurance_waiting_time.csv?

Exercise 2: An two wheeler manufacturing company wants to study the effect of four engine tuning techniques on the mileage. The data collected is given in Mileage.csv file. Test whether the tuning techniques impacts the mileage?

CROSS TABULATION

CROSS TABULATION

- An approach to summarize and identify the relation between two or more variables or parameters
- Describes two variables simultaneously
- Expressed as two way table
- Variables need to be categorical or grouped

Input or Process Variable	Output Variable				
	Very Good	Good	Average	Below Average	Poor
0 – 3					
3 - 6					
6 - 12					

CROSS TABULATION

Indian Statistical Institute

Example: An ITeS company has collected the data from 116 employees on their response to the appraisal results, their gender, location and vintage. The data is given in Appraisal_Response file .

1. Does male and female differ in their response to appraisal?
2. Does employees in different locations differ in their response to appraisal?
3. Does employees with different vintages differ in their response to appraisal?

175

CROSS TABULATION

Indian Statistical Institute

a. Reading the file and converting variables to factors

```
mydata = Appraisal_Response
response = mydata$Response
gender = mydata$Gender
location = mydata$Location
experience = mydata$Experience
```

176

CROSS TABULATION

b. Constructing cross tabulation of Gender vs. Response to appraisal

```
mytable = table(gender, response)
mytable
```

Gender	Appraisal Response			
	Disappointed	No Comments	Happy	Total
Female	12	12	36	60
Male	30	12	14	56
Total	42	24	50	116

177

CROSS TABULATION

c. Constructing cross tabulation of Gender vs. Appraisal Response – %

```
prop.table(mytable)*100
```

Gender	Appraisal Response		
	Disappointed	No Comments	Happy
Female	10.34	10.34	31.03
Male	25.86	10.34	12.07

178

CROSS TABULATION

Indian Statistical Institute

d. Constructing cross tabulation of Gender vs. Appraisal Response – row %

prop.table(mytable,1)*100

Gender	Appraisal Response		
	Disappointed	No Comments	Happy
Female	20.00	20.00	60.00
Male	53.57	21.43	25.00

179

CROSS TABULATION

Indian Statistical Institute

d. Constructing cross tabulation of Gender vs. Appraisal response – column %

> prop.table(mytable,2)*100

Gender	Appraisal Response		
	Disappointed	No Comments	Happy
Female	28.57	50.00	72.00
Male	71.43	50.00	28.00

180

CHI SQUARE TEST

CHI SQUARE TEST

Objective:

To test whether two variables are related or not

To check whether a metric is depends on another metric

Usage:

When both the variables (x & y) need to be categorical (grouped)

H0: Relation between x & y = 0 or x and y are independent

H1: Relation between x & y $\neq$ 0 or x and y are not independent

If p value < 0.05, then H0 is rejected

CHI SQUARE TEST

Exercise:

A project is undertaken to improve the CSat score of transaction processing. Based on brainstorming, the project team suspects that lack of experience is a cause of low CSat score.

The following data was collected. Analyze the data and verify whether CSat score dependents on experience

Experience (Months)	CSat Score				
	VD	D	N	S	VS
0 – 3	50	40	30	10	10
3- 6	5	30	50	35	7
6 - 9	6	7	30	40	50

Note: Table gives the count of CSat score of very dissatisfied to very satisfied for agents belonging to three different experience groups

CHI SQUARE TEST

Exercise:

Step 1: Calculate the row and column sum

Experience (Months)	CSat Score					
	VD	D	N	S	VS	Row Sum
0 – 3	50	40	30	10	10	140
3 - 6	5	30	50	35	7	127
6 - 9	6	7	30	40	50	133
Col Sum	61	77	110	85	67	400

CHI SQUARE TEST

Exercise:

Step 2: Calculate expected count for each cell

Expected count of CSat score VD for group 0 – 3 months experience

= Expected count of cell (1,1) = (Row 1 sum x Column 1 sum) / Total

= (140 x 61) / 400 = 21.4

Table of expected count (the count expected if variables are not related)

Experience (Months)	CSat Score					
	VD	D	N	S	VS	Row Sum
0 – 3	21.4	27	38.5	29.8	23.5	140
3 - 6	19.4	24.4	34.9	27	21.3	127
6 - 9	20.3	25.6	36.6	28.3	22.3	133
Col Sum	61	77	110	85	67	400

185

CHI SQUARE TEST

Exercise:

Step 3: Take difference between observed count and expected count

For cell (1,1)

observed Count = 50

expected Count = 21.4

difference = 28.7

Table of observed count – expected count

Experience (Months)	CSat Score				
	VD	D	N	S	VS
0 – 3	28.7	13.1	-8.5	-20	-13
3 - 6	-14.4	5.55	15.1	8.01	-14
6 - 9	-14.3	-19	-6.6	11.7	27.7

186

CHI SQUARE TEST

Exercise:

Step 4: Calculate (observed - expected)^2 / expected for each cell

Table of (observed - expected)^2 / expected

Experience (Months)	CStat Score				
	VD	D	N	S	VS
0 – 3	38.45	6.32	1.88	13.11	7.71
3 - 6	10.66	1.26	6.51	2.38	9.58
6- 9	10.06	13.52	1.18	4.87	34.50

187

CHI SQUARE TEST

Exercise:

Step 5: Calculate Chi Square = Sum of all ((observed - expected)^2 / expected)

Chi Square calculated = 38.45 + 6.32 + - - - + 34.5

Chi Square Calculated χ^2= 161.98

If variables are not related then χ^2 will be close to 0

Step 6: Calculate p value

P value = chidist(chi Sq, df)

= chidist(161.98,8)

= 0.00

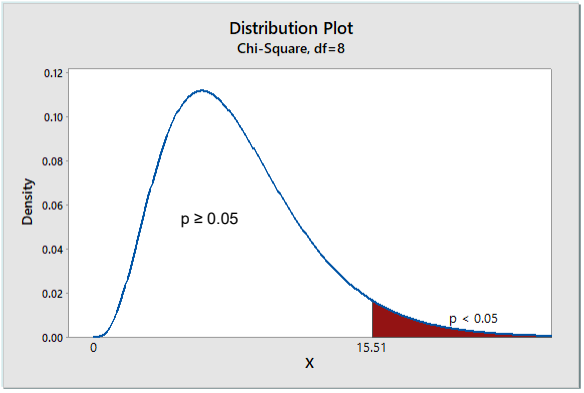
Conclusion:

Since p value 0.00 < 0.05, Csat score depends on experience or the variables are related

188

CHI SQUARE TEST

Exercise:



Conclusion:

Since p value 0.00 < 0.05, Csat score depends on experience or the variables are related

CHI SQUARE TEST

Issues:

- Chi square test only shows whether two variables are independent or not
- Degree of association will not be known

Measures of Strength of relationship:

1. Phi (ϕ) Coefficient

$$\phi = \sqrt{(\chi^2 / n)}$$

Only for 2 x2 tables

2. Cramer V = $\sqrt{(\phi^2 / (\min (\text{rows} - 1) , (\text{cols} - 1)))}$

Phi & Cramer V varies from 0 to 1, higher the value better the strength of relation

CHI SQUARE TEST

Phi Coefficient = sqrt(161.98 / 400) = 0.64

Cramer V:

Rows – 1 = 2

Columns - 1 = 4

Cramer V = $\sqrt{}$(0.64^2/ 2) = 0.4499 = 44.99%

CHI SQUARE TEST

Example: An ITeS company has collected the data from 116 employees on their response to the appraisal results, their gender, location and vintage. The data is given in Appraisal_Response file .

1. Is the response to appraisal vary with gender?
2. Is the appraisal response same across locations?
3. Is the response to appraisal different for employees with differnt vintage?

CROSS TABULATION

a. Reading the file and converting variables to factors

```
mydata = Appraisal_Response
response = mydata$Response
gender = mydata$Gender
location = mydata$Location
experience = mydata$Experience
```

193

CROSS TABULATION

b. Constructing cross tabulation of Gender vs. Response to appraisal

```
mytable = table(gender, response)
mytable
```

Gender	Appraisal Response		
	Disappointed	No Comments	Happy
Female	12	12	36
Male	30	12	14

194

CHI SQUARE TEST

Indian Statistical Institute

c. Chi Square test of independence - Gender vs. Response

> chisq.test(mytable)

Statistic	Value
Chi Square	17.277
df	2
p_value	0.00018
Critical Value	5.99

195

CHI SQUARE TEST

Indian Statistical Institute

Fisher's Exact test

When one or more of expected frequencies are less than 5

d. Fisher's exact test of independence - Gender vs. Response

> fisher.test(mytable)

Statistic	Value
p_value	0.00014

196

CHI SQUARE TEST

e. Measures of Association - Gender vs. Usage

> library(vcd)

> assocstats(mytable)

	Chi Square	df	p - value
Likelihood Ratio	17.851	2	0.00013
Pearson	17.277	2	0.00018

Statistics	Value
Phi-Coefficient	NA
Contingency Coefficient	0.36
Cramer's V	0.386

CORRELATION
&
REGRESSION

Indian Statistical Institute

CORRELATION & REGRESSION

Correlation:

Correlation analysis is a technique to identify the relationship between two variables.

Type and degree of relationship between two variables.

199

Indian Statistical Institute

CORRELATION & REGRESSION

Correlation: Usage

Explore the relationship between the output characteristic and input or process variable.

Output variable : y : Dependent variable

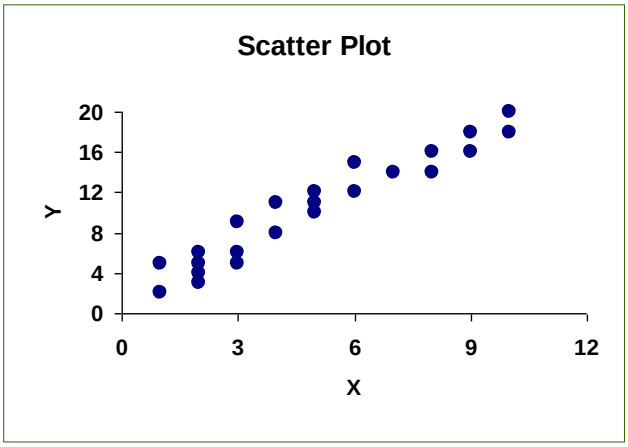
Input / Process variable : x : Independent variable

200

Indian Statistical Institute

CORRELATION & REGRESSION

Positive Correlation: y increases as x increases & vice versa

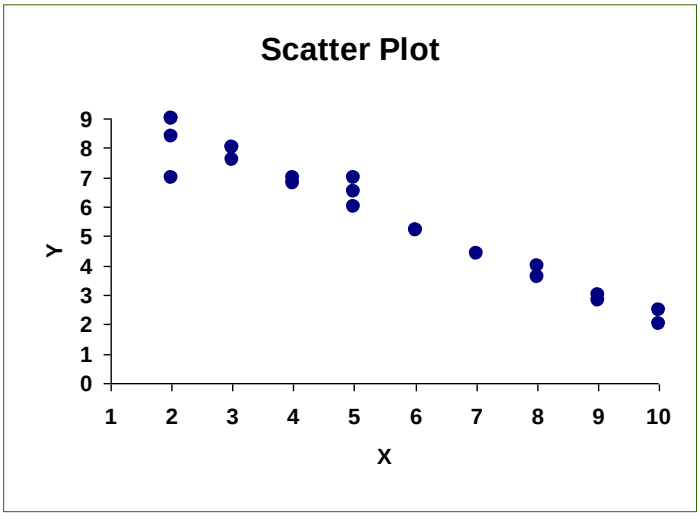


201

Indian Statistical Institute

CORRELATION & REGRESSION

Negative Correlation: y decreases as x increases & vice versa

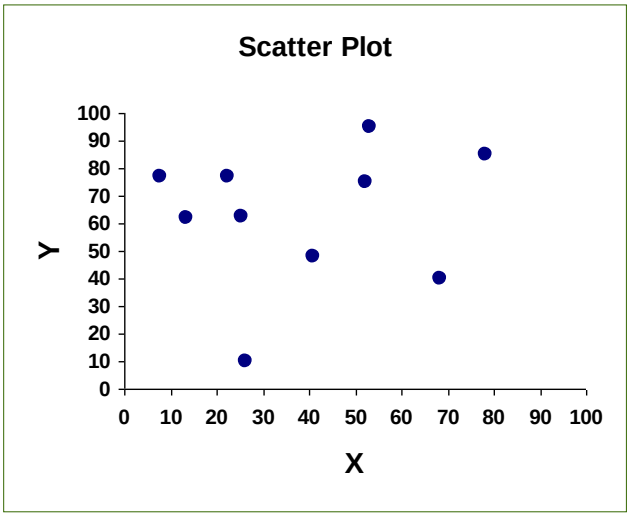


202

Indian Statistical Institute

CORRELATION & REGRESSION

No Correlation: Random Distribution of points

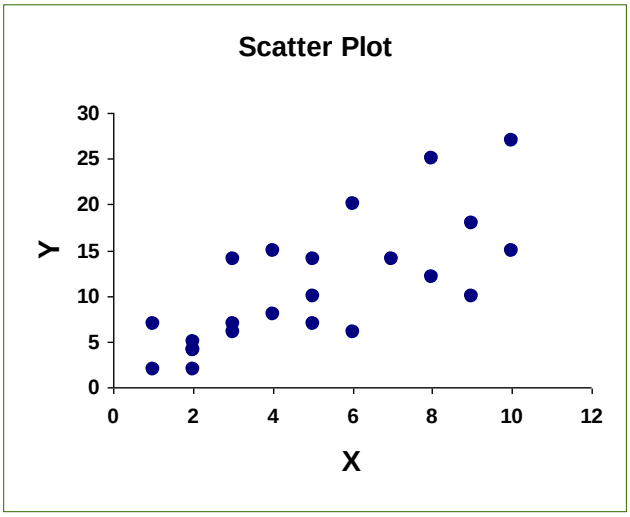


203

Indian Statistical Institute

CORRELATION & REGRESSION

Is there any correlation ?



204

Indian Statistical Institute

CORRELATION & REGRESSION

Measure of Correlation: Coefficient of Correlation

Symbol : r

Range : -1 to 1

Sign : Type of correlation

Value : Degree of correlation

Examples:

r = 0.6 , 60 % positive correlation

r = -0.82, 82% negative correlation

r = 0, No correlation

205

Indian Statistical Institute

CORRELATION & REGRESSION

Coefficient of Correlation Computation: Positive Correlation

Collect data on x and y: When x is low, y is also low & vice versa

x	y
2	5
3	7
1	3
5	11
6	12
7	15

206

CORRELATION & REGRESSION

Coefficient of Correlation Computation : Positive Correlation

Calculate Mean of x & y values

SL No.	x	y
1	2	5
2	3	7
3	1	3
4	5	11
5	6	12
6	7	15
Mean	4	8.83

CORRELATION & REGRESSION

Coefficient of Correlation Computation : Positive Correlation

Take x – Mean x and y – Mean y

SL No.	x – Mean x	y – Mean y
1	-2	-3.83
2	-1	-1.83
3	-3	-5.83
4	1	2.17
5	2	3.17
6	3	6.17

Conclusion:

Low values will become negative & high values will become positive

CORRELATION & REGRESSION

Coefficient of Correlation Computation : Positive Correlation

Generally when x values are negative, y values are also negative & vice versa

SL No.	x – Mean x	y – Mean y
1	-2	-3.83
2	-1	-1.83
3	-3	-5.83
4	1	2.17
5	2	3.17
6	3	6.17

209

CORRELATION & REGRESSION

Coefficient of Correlation Computation : Positive Correlation

Then

Product of x & y values will be generally positive

SL No.	x – Mean x	y – Mean y	Product
1	-2	-3.83	7.66
2	-1	-1.83	1.83
3	-3	-5.83	17.49
4	1	2.17	2.17
5	2	3.17	6.34
6	3	6.17	18.51
		Sum = Sxy	54

210

CORRELATION & REGRESSION

Coefficient of Correlation Computation : Positive Correlation

Sum of Product of x & y values (Sxy) will be positive

SL No.	x – Mean x	y – Mean y	Product
1	-2	-3.83	7.66
2	-1	-1.83	1.83
3	-3	-5.83	17.49
4	1	2.17	2.17
5	2	3.17	6.34
6	3	6.17	18.51
		Sum = Sxy	54

211

CORRELATION & REGRESSION

Coefficient of Correlation Computation : Negative Correlation

Collect data on x and y: When x is low then y will be high & vice versa

x	y
2	12
3	11
1	15
5	7
6	5
7	3

212

CORRELATION & REGRESSION

Coefficient of Correlation Computation : Negative Correlation

Calculate Mean of x & y values

SL No.	x	y
1	2	12
2	3	11
3	1	15
4	5	7
5	6	5
6	7	3
Mean	4	8.83

213

CORRELATION & REGRESSION

Coefficient of Correlation Computation : Negative Correlation

Take x – Mean x and y – Mean y

SL No.	x – Mean x	y – Mean y
1	-2	3.67
2	-1	2.67
3	-3	6.67
4	1	-1.33
5	2	-3.33
6	3	-5.33

Conclusion:

Low values will become negative & high values will become positive

214

CORRELATION & REGRESSION

Coefficient of Correlation Computation : Negative Correlation

Generally when x values are negative, y values are positive & vice versa

SL No.	x – Mean x	y – Mean y
1	-2	3.67
2	-1	2.67
3	-3	6.67
4	1	-1.33
5	2	-3.33
6	3	-5.33

215

CORRELATION & REGRESSION

Coefficient of Correlation Computation : Negative Correlation

Then

Product of x & y values will be generally negative

SL No.	x – Mean x	y – Mean y	Product
1	-2	3.67	-7.34
2	-1	2.67	-2.67
3	-3	6.67	-20.01
4	1	-1.33	-1.33
5	2	-3.33	-6.66
6	3	-5.33	-15.99
		Sum = Sxy	- 54

216

CORRELATION & REGRESSION

Coefficient of Correlation Computation : Negative Correlation

Sum of Product of x & y values Sxy will be negative

SL No.	x – Mean x	y – Mean y	Product
1	-2	3.67	-7.34
2	-1	2.67	-2.67
3	-3	6.67	-20.01
4	1	-1.33	-1.33
5	2	-3.33	-6.66
6	3	-5.33	-15.99
		Sum = Sxy	- 54

217

CORRELATION & REGRESSION

Coefficient of Correlation Computation :

In Short

If correlation is positive

Sxy will be positive

If correlation is negative

Sxy will be negative

218

CORRELATION & REGRESSION

Coefficient of Correlation Computation :

Sxy is divided by $\sqrt{}$ (Sxx.Syy)

Sxy = Σ(x-Mean x)(y-Mean y)

Sxx = Σ(x-Mean x)^2

Syy = Σ(y-Mean y)^2

Correlation Coefficient r = Sxy / $\sqrt{}$ (Sxx.Syy)

CORRELATION & REGRESSION

Coefficient of Correlation Computation :

SL No.	x – Mean x	y – Mean y	Product	(x – Mean x)^2	(y – Mean y)^2
1	-2	3.67	-7.34	4	14.6689
2	-1	2.67	-2.67	1	3.3489
3	-3	6.67	-20.01	9	33.9889
4	1	-1.33	-1.33	1	4.7089
5	2	-3.33	-6.66	4	10.0489
6	3	-5.33	-15.99	9	38.0689
Sum			Sxy: -54	Sxx: 28	Syy:104.83

r = Sxy / $\sqrt{}$Sxx.Syy = -54 / $\sqrt{}$(28 x 104.83) = -0.9967

Indian Statistical Institute

CORRELATION & REGRESSION

Correlation Coefficients:

1. Spearman's rho (ρ)
2. Kendall's Tau (τ)

Varies from -1 to +1

Close to -1 indicate negative correlation

Close to +1 indicate positive correlation

Close to 0 means no correlation

Generally used for non normal or non measurable data

221

Indian Statistical Institute

CORRELATION & REGRESSION

Exercise: The data on vapor pressure of water at various temperatures are given in Correlation.csv file.

1. Construct the scatter plot and interpret?
2. Compute the correlation coefficient?
3. Test whether the correlation coefficient is equal to 0?
4. Compute 95% confidence interval on correlation coefficient?

222

CORRELATION & REGRESSION

Exercise: The data on vapor pressure of water at various temperatures are given in Correlation.csv file.

1. Reading the data and variables

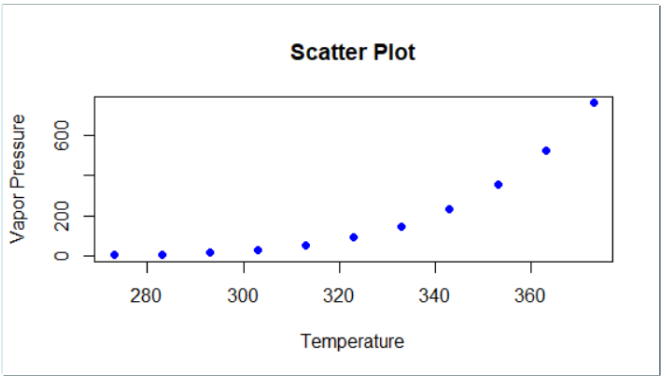
```
> mydata = Correlation
> Temp = mydata$Temperature
> Pressure = mydata$Vapor.Pressure
```


CORRELATION & REGRESSION

Exercise: The data on vapor pressure of water at various temperatures are given in Correlation.csv file.

2. Constructing Scatter plot

```
> plot(Temp, Pressure, main = "Scatter Plot", xlab = "Temperature", ylab = "Vapor Pressure", col = "blue", pch = 19))
```



CORRELATION & REGRESSION

Exercise: The data on vapor pressure of water at various temperatures are given in Correlation.csv file.

Computing correlation coefficient

```
> cor(Temp, Pressure)
```

Statistics	Value
r	0.893

CORRELATION & REGRESSION

Exercise: The data on vapor pressure of water at various temperatures are given in Correlation.csv file.

Testing correlation coefficient = 0 and Computing correlation coefficient

```
cor.test(Temp, Pressure)
```

Statistics	Value
t statistic	5.9617
p value	0.0002122

95 % Lower Confidence Limit	95 % Upper Confidence Limit
0.6321	0.9722

MULTIPLE REGRESSION ANALYSIS

CORRELATION & REGRESSION

Regression

Correlation helps

To check whether two variables are related

If related

Identify the type & degree of relationship

Indian Statistical Institute

CORRELATION & REGRESSION

Regression

Regression helps

- To identify the exact form of the relationship
- To model output in terms of input or process variables

Examples:

Expected (Yield) = 5 + 3 x Time - 2 x Temperature

229

Indian Statistical Institute

CORRELATION & REGRESSION

Simple Linear Regression Illustration

Output variable is modeled in terms of only one variable

x	y
2	7
1	4
5	16
4	13
3	10
6	19

Regression Model

$y = 1 + 3x$

230

CORRELATION & REGRESSION

Simple Linear Regression

General Form:

$$y = a + bx + \varepsilon$$

where

a: intercept (the value of y when x is equal to 0)

b: slope (indicates the amount of change in y with every unit change in x)

CORRELATION & REGRESSION

Simple Linear Regression: Parameter Estimation

Model: $y = a + bx + \varepsilon$

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

$$\hat{b} = S_{xy} / S_{xx}$$

Test for Significance (Testing b = 0 or not) of relation between x & y

H0: $b = 0$

H1: $b \neq 0$

Test Statistic $t_0 = (\hat{b} - 0)/se(\hat{b})$

If p value < 0.05, then H0 is rejected & y can be modeled with x

CORRELATION & REGRESSION

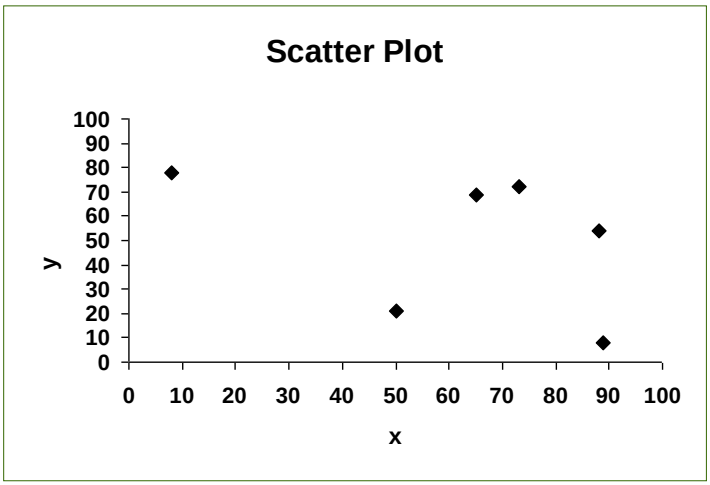
Regression illustration: Issues

x	y
65	69
8	78
89	8
88	21
50	24
73	72

233

CORRELATION & REGRESSION

Regression Model $y = 76.32 - 0.42x + \varepsilon$



234

CORRELATION & REGRESSION

Regression: Issues

For any set of data,

a & b can be calculated

Regression model $y = a + bx + \varepsilon$ can be build

But all the models may not be useful

CORRELATION & REGRESSION

Coefficient of Regression: Measure of degree of Relationship

Symbol : R^2

$R^2 = SS_R / Syy = b.Sxy / Syy$

$SS_R = \Sigma(y_{predicted} - \text{Mean } y)^2$

$Syy = \Sigma(y_{actual} - \text{Mean } y)^2$

R^2 : amount variation in y explained by x

Range of R^2 : 0 to 1

If $R^2 \geq 0.6$, the model is reasonably good

CORRELATION & REGRESSION

Coefficient of Regression: Testing the significance of Regression

Regression ANOVA

Model	SS	df	MS	F	p value
Regression	SS_R				
Residual	Syy – SS_R				
Total	Syy				

If p value < 0.05, then the regression model is significant

REGRESSION ANALYSIS

Multiple Linear Regression

To model output variable y in terms of two or more variables.

General Form:

$$y = a + b_1x_1 + b_2x_2 + - - - + b_kx_k + \varepsilon$$

Two variable case:

$$y = a + b_1x_1 + b_2x_2 + \varepsilon$$

Where

a: intercept (the predicted value of y when all x's are zero)

b_j: slope (the amount change in y for unit change in x_j keeping all other x's constant, j = 1,2,---,k)

Indian Statistical Institute

REGRESSION ANALYSIS

Exercise : The effect of temperature and reaction time affects the % yield. The data collected in given in the Mult-Reg_Yield file. Develop a model for % yield in terms of temperature and time?

Step 1: Read data

```
> mydata = mydata[,2:4]
> attach(mydata)
> temp = Temperature
 > time = Time
> yield = X.Yield
```

239

Indian Statistical Institute

REGRESSION ANALYSIS

Exercise : The effect of temperature and reaction time affects the % yield. The data collected in given in the Mult-Reg_Yield file. Develop a model for % yield in terms of temperature and time?

Step 1: Correlation Analysis

```
> cor(mydata)
```

Attribute	Time	Temperature	% Yield
Time	1.00	-0.01	0.90
Temperature	-0.01	1.00	-0.05
% Yield	0.90	-0.05	1.00

Correlation between xs & y should be high

Correlation between xs should be low

240

Indian Statistical Institute

REGRESSION ANALYSIS

Exercise : The effect of temperature and reaction time affects the % yield. The data collected in given in the Mult-Reg_Yield file. Develop a model for % yield in terms of temperature and time?

Step 1: Correlation Analysis – Scatter Plot

```
# Scatter plot
library(caret)
y = mydata$Yield
x = mydata[,1:2]

featurePlot(x, y, plot = "scatter", pch = 19)
```

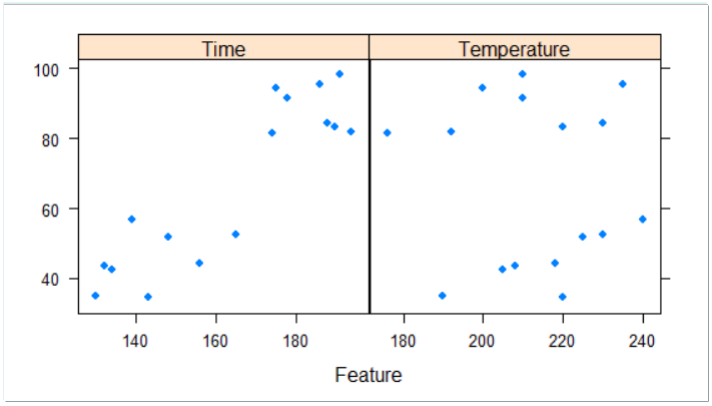
241

Indian Statistical Institute

REGRESSION ANALYSIS

Exercise : The effect of temperature and reaction time affects the % yield. The data collected in given in the Mult-Reg_Yield file. Develop a model for % yield in terms of temperature and time?

Step 1: Correlation Analysis – Scatter Plot



242

Indian Statistical Institute

REGRESSION ANALYSIS

Step 2: Regression Output

```
# Regression Model
attach(mydata)
mymodel = lm(Yield ~ Time + Temperature)
summary(mymodel)
```

Statistic	Value	Criteria
R Square	0.8064	≥ 0.6
Adjusted R Square	0.7766	≥ 0.6

243

Indian Statistical Institute

REGRESSION ANALYSIS

Step 2: Regression Output

Regression ANOVA

Model	SS	df	MS	F	p value
Regression	6797.063	2	3398.531	27.07	0.0000
Residual	1632.08138	13	125.5447		
Total	8429.14438	15			

Criteria: P value < 0.05

244

REGRESSION ANALYSIS

Step 2: Regression Output – Identify the model

Attribute	Coefficient	Std. Error	t Statistic	p value
Time	0.9061	0.12337	7.344	0.0000
Temperature	-0.0642	0.16391	-0.392	0.702
Intercept	-67.8844	40.58652	-1.67	0.118

Interpretation: Only time is related to yield as p value < 0.05

REGRESSION ANALYSIS

Step 2: Regression Output – Identify the model

Attribute	Coefficient	Std. Error	t Statistic	p value
Time	0.9065	0.1196	7.580	0.0000
Intercept	-81.6205	19.7906	-4.124	0.00103

Model Yield= 0.9065 x Time - 81.621

Indian Statistical Institute

REGRESSION ANALYSIS

Step 3: Residual Analysis

```
> pred = fitted(mymodel, mydata)
> res = residuals(mymodel)
> cbind(yield, pred, res)
```

SL No	Conversion	Fitted	Residuals
1	49	45.8382	3.1618062
2	50.2	45.8382	4.3618062
3	50.5	46.0101	4.4899396
4	48.5	45.4945	3.0055393
5	47.5	45.3226	2.1774058
6	44.5	45.8382	-1.3381938
7	28	36.2137	-8.2136666
8	31.5	36.9011	-5.4011328
9	34.5	38.9635	-4.4635315
10	35	41.0259	-6.0259302
11	38	38.2761	-0.2760653
12	38.5	35.8699	2.6300665
13	15	21.0894	-6.0894095
14	17	16.2772	0.7228541
15	20.5	18.3395	2.1604554
16	29.5	20.4019	9.0980567

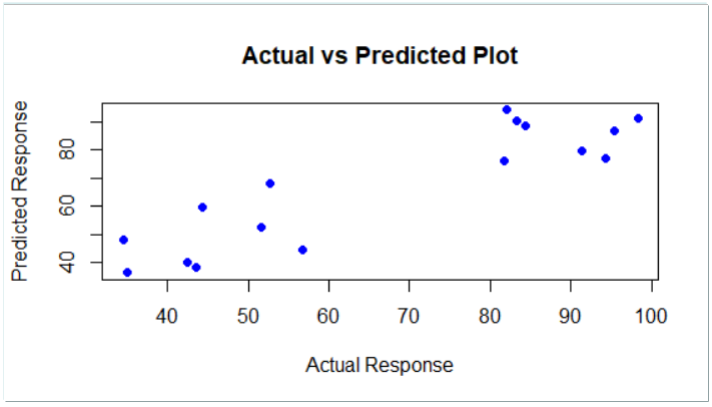
247

Indian Statistical Institute

REGRESSION ANALYSIS

Step 3: Residual Analysis – Actual Vs Fitted

```
> plot(Yield, pred)
```



Note: There need to be strong positive correlation between actual and fitted response

248

REGRESSION ANALYSIS

Step 3: Residual Analysis:

```
> qqnorm(res)
> qqline(res)
> shapiro.test(res)
```

Shapiro-Wilk normality Test: Yield data	
W	p value
0.94219	0.3767

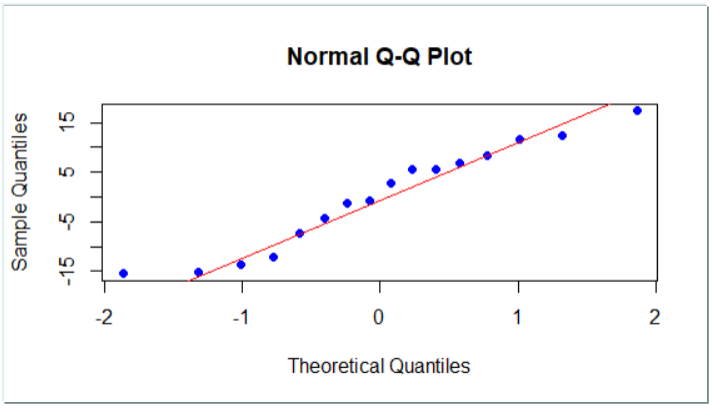
```
> mse = mean(res^2)
> rmse = sqrt(mse)
```

Statistic	Value
MSE	102.0051
RMSE	10.09976

249

REGRESSION ANALYSIS

Step 3: Residual Analysis: Normal Q-Q plot of residuals



250

REGRESSION ANALYSIS

6: Outlier test

Observations with Bonferonni p – value < 0.05 are potential outliers

```
> library(car)
> outlierTest(mymodel)
```

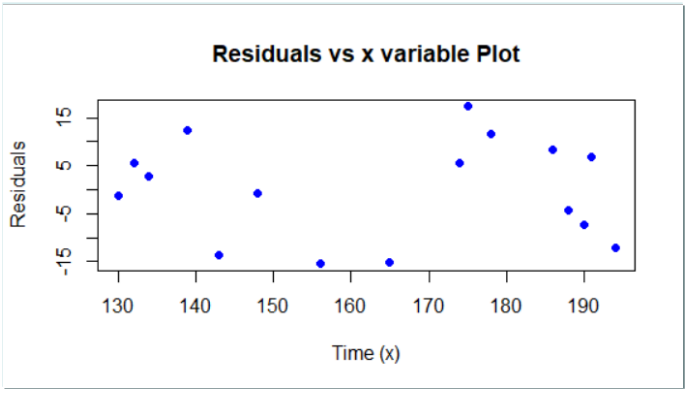
Observation	Studentized Residual	Bonferonni p value
11	1.781515	NA

REGRESSION ANALYSIS

7: Model adequacy check

Residuals Vs Independent variables

```
> plot(Time, res)
```



Note: There should not be any pattern or trend, the points should be distributed randomly

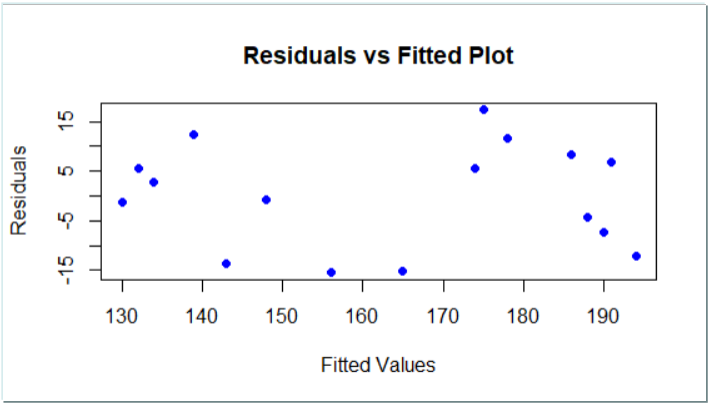
Indian Statistical Institute

REGRESSION ANALYSIS

7: Model adequacy check

Residuals Vs Fitted

```
> plot(pred, res)
```



Note: There should not be any pattern or trend, the points should be distributed randomly

253

Indian Statistical Institute

REGRESSION ANALYSIS

7: Model validation

```
> mse = mean(res^2)
> rmse = sqrt(mse)
```

Statistics	Value
Mean Square Error (MSE)	103.2084
Root Mean Square Error (RMSE)	10.15915

254

REGRESSION ANALYSIS

7: Leave One Out Cross Validation (LOOCV)

```
> library(boot)
> mymodel = glm(Yield ~ Time)
> myvalidation = cv.glm(mydata, mymodel)
> loocv_mse = mycv$delta[1]
> loocv_mse
> loocv_rmse = sqrt(loocv_mse)
> loocv_rmse
```

	Original	LOOCV
MSE	103.2084	128.8541
RMSE	10.15915	11.35139

REGRESSION ANALYSIS

7: k fold Cross Validation (LOOCV)

```
# k = 4 fold cross validation
set.seed(1)
mymodel = glm(Yield ~ Time)
mycv = cv.glm(mydata, mymodel, K = 4)
k_fold_cv_mse = mycv$delta[1]
k_fold_cv_rmse = sqrt(k_fold_cv_mse)
k_fold_cv_mse
k_fold_cv_rmse
```

	Original	LOOCV	K – fold CV
MSE	103.2084	128.8541	178.0298
RMSE	10.15915	11.35139	13.34278

REGRESSION ANALYSIS

Example 2 : The effect of temperature, time and kappa number of pulp affects the % conversion of UB pulp to Cl_2 pulp. inspection. The data collected in given in the Mult_Reg_Conversion file. Develop a model for % conversion in terms of exploratory variables?

REGRESSION ANALYSIS

Step 1: Correlation Analysis

	Temperature	Time	Kappa #	% Conversion
Temperature	1.00	-0.96	0.22	0.95
Time	-0.96	1.00	-0.24	-0.91
Kappa #	0.22	-0.24	1.00	0.37
% Conversion	0.95	-0.91	0.37	1.00

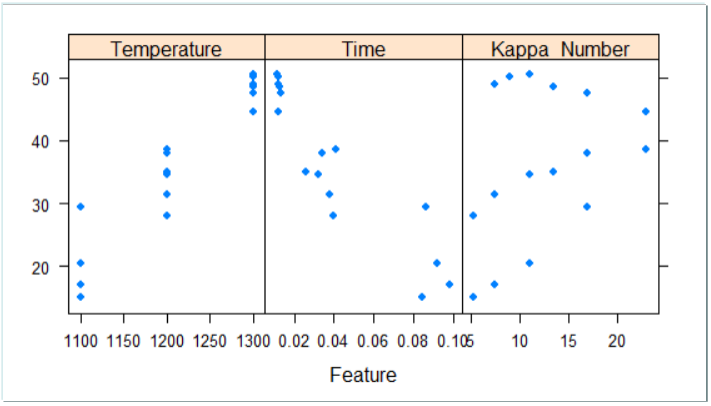
Interpretation

High Correlation between % Conversion and Temperature & Time

High Correlation between Temperature & Time - Multicollinearity

REGRESSION ANALYSIS

Step 1: Correlation Analysis: Scatter Plot



259

REGRESSION ANALYSIS

Measure for Multicollinearity

Variance Inflation Factor (VIF)

Measures the correlation (linear association) between each x variable with other x's

$VIF_i = 1/(1- R_i^2)$

Where R_i is the coefficient for regressing x_i on other x's

Criteria: VIF < 5

260

Indian Statistical Institute

REGRESSION ANALYSIS

Regression Output

Statistic	Value	Criteria
Adjusted R Square	0.899	> 0.6

Regression ANOVA

Model	SS	df	MS	F	p value
Regression	1953.419	3	651.140	45.885	0.0000
Residual	170.290	12	14.191		
Total	2123.709	15			

261

Indian Statistical Institute

REGRESSION ANALYSIS

Regression Output

	Coeff	Std. Error	t	p value
Constant	-121.27	55.43571	-2.19	0.0492
Temperature	0.12685	0.04218	3.007	0.0109
Time	-19.0217	107.92824	-0.18	0.863
Kappa #	0.34816	0.17702	1.967	0.0728

Variance-inflation factors (VIF)

```
> library(car)
> vif(mymodel)
```

x	VIF
Temperature	12.23
Time	12.33
Kappa #	1.062

262

REGRESSION ANALYSIS

Tackling Multicollinearity:

1. Remove one or more of highly correlated independent variable
2. Principal Component Regression
3. Partial Least Square Regression
4. Ridge Regression

REGRESSION ANALYSIS

Tackling Multicollinearity:

Method 1: Removing highly correlated variable – Stepwise Regression

Approach

- A null model is developed without any predictor variable x. In null model, the predicted value will be the overall mean of y
- Then predictor variables x's are added to the model sequentially
- After adding each new variable, the method also remove any variable that no longer provide an improvement in the model fit
- Finally the best model is identified as the one which minimizes Akaike information criterion *(*AIC)

$$AIC = \frac{1}{n\hat{\sigma}^2}(RSS + 2d\hat{\sigma}^2)$$

REGRESSION ANALYSIS

Tackling Multicollinearity:

Method 1: Removing highly correlated variable – Stepwise Regression

Akaike information criterion (AIC)

$$AIC = \frac{1}{n\hat{\sigma}^2}(RSS + 2d\hat{\sigma}^2)$$

n: number of observations

$\hat{\sigma}^2$: estimate of error or residual variance

d: number of x variables included in the model

RSS: Residual sum of squares

REGRESSION ANALYSIS

Tackling Multicollinearity:

Method 1: Removing highly correlated variable – Stepwise Regression

R code

```
> library(MASS)
> mymodel = lm(X..Conversion ~ Temperature + Time + Kappa.number)
> step =stepAIC(mymodel, direction = "both")
```

Step	x's in the model	AIC
1	Temperature, Time & Kappa Number	45.8
2	Temperature & Kappa Number	43.9

REGRESSION ANALYSIS

Tackling Multi collinearity:

Method 1: Stepwise Regression

Attribute	Coefficient	Std. Error	t Statistic	p value
Temperature	0.13396	0.01191	11.250	0.0000
Kappa #	0.35106	0.16955	2.071	0.0589
Intercept	-130.68986	14.14571	-9.239	0.0000

% Conversion = 0.13396 * Temperature + 0.35106 * Kappa # - 130.68986

Variance-inflation factors (VIF)

x	VIF
Temperature	1.0526
Kappa #	1.0526

Statistic	Value
R^2	0.9196
Adjusted R^2	0.9072

267

REGRESSION ANALYSIS

Regression with interaction:

```
mymodel = lm(Conversion ~ Temperature + Kappa_Number + Temperature :
            Kappa_Number)
Summary(mymodel)
```

Statistic	Value
R^2	0.9848
Adjusted R^2	0.981

	Estimate	Std. Error	t value	p value
(Intercept)	-238.40	16.33	-14.599	0.00
Temperature	0.22	0.01	16.473	0.00
Kappa_Number	9.62	1.30	7.429	0.00
Temperature:Kappa_Number	-0.01	0.00	-7.17	0.00

268

REGRESSION ANALYSIS

Residual Analysis:

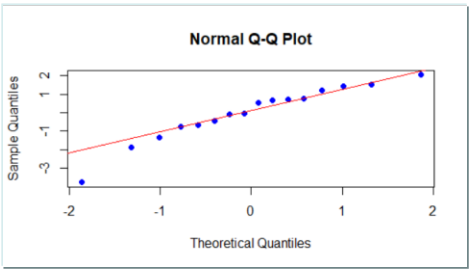
```
# Residual analysis
pred = predict(mymodel, mydata, interval = "prediction")
res = residuals(mymodel)
myresult = cbind(mydata, pred, res)
myresult =round(myresult,3)
```

SL No	Temperature	Time	Kappa_Number	Conversion	fit	lwr	upr	res
1	1300	0.012	7.5	49	49.424	45.335	53.514	-0.424
2	1300	0.012	9	50.2	49.011	45.041	52.981	1.189
3	1300	0.012	11	50.5	48.459	44.596	52.322	2.041
4	1300	0.013	13.5	48.5	47.77	43.948	51.591	0.73
5	1300	0.014	17	47.5	46.804	42.867	50.742	0.696
6	1300	0.012	23	44.5	45.15	40.599	49.7	-0.65
7	1200	0.04	5.3	28	31.765	27.898	35.632	-3.765
8	1200	0.038	7.5	31.5	32.832	29.059	36.605	-1.332
9	1200	0.032	11	34.5	34.53	30.832	38.229	-0.03
10	1200	0.026	13.5	35	35.743	32.038	39.448	-0.743
11	1200	0.034	17	38	37.441	33.645	41.238	0.559
12	1200	0.041	23	38.5	40.353	36.195	44.511	-1.853
13	1100	0.084	5.3	15	13.498	9.211	17.786	1.502
14	1100	0.098	7.5	17	16.24	12.192	20.288	0.76
15	1100	0.092	11	20.5	20.601	16.697	24.506	-0.101
16	1100	0.086	17	29.5	28.078	23.709	32.448	1.422

269

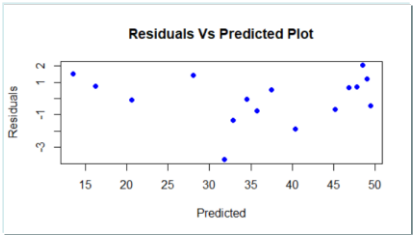
REGRESSION ANALYSIS

Residual Plots:



Normality Test - Residuals

Statistic	Value
w	0.9318
p_value	0.2603



270

REGRESSION ANALYSIS

Cross Validation:

```
mse = mean(res^2)
rmse = sqrt(mse)

# LOOCV
library(boot)
mymodel = glm(Conversion ~ Temperature + Kappa_Number +
Temperature:Kappa_Number)
mycv = cv.glm(mydata, mymodel)
mse = mycv$delta[1]
rmse = sqrt(mse)

# k-fold CV
set.seed(1)
mymodel = glm(Conversion ~ Temperature + Kappa_Number +
Temperature:Kappa_Number)
mycv = cv.glm(mydata, mymodel, K = 4)
mse = mycv$delta[1]
rmse = sqrt(mse)
```


REGRESSION ANALYSIS

Cross Validation:

Statistic	Training	LOOCV	k=4 fold CV
MSE	2.0192	3.8421	5.6391
RMSE	1.4210	1.9601	2.3747

REGRESSION ANALYSIS

Regression with dummy variables

When x's are not numeric but nominal

Each nominal or categorical variable is converted into dummy variables

Dummy variables takes values 0 or 1

Number of dummy variable for one x variable is equal to number of distinct values of that variable - 1

Example: A study was conducted by an IT company to develop a model to estimate sprint productivity of agile projects in telecom vertical. The explonatory variables chosen are developer skill, review type and code reuse. Data was collected from 34 projects and is given in Agile_Productivity file.?

REGRESSION ANALYSIS

Regression with dummy variables

Variable		Dummy
Review Type	Code	$Review_{Peer}$
Fagan	1	0
Peer	2	1

Variable		Dummy	
Experience	Code	$Experience_{Experienced}$	$Experience_{Master}$
Fresher	1	0	0
Experienced	2	1	0
Master	3	0	1

REGRESSION ANALYSIS

Regression with dummy variables

Read the fie and variables

```
mydata =  Agile_Productivity
skill = mydata$Developer_Skill
review =mydata$Review_Type
reuse = mydata$Reuse
SP = mydata$Sprint_Productivty
```

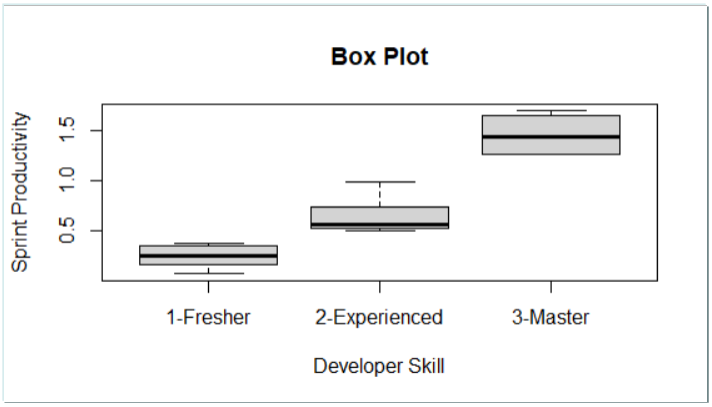
Converting categorical x's to factors

```
skill = factor(skill)
review = factor(review)
contrasts(skill)
contrasts(review)
```


REGRESSION ANALYSIS

Regression with dummy variables

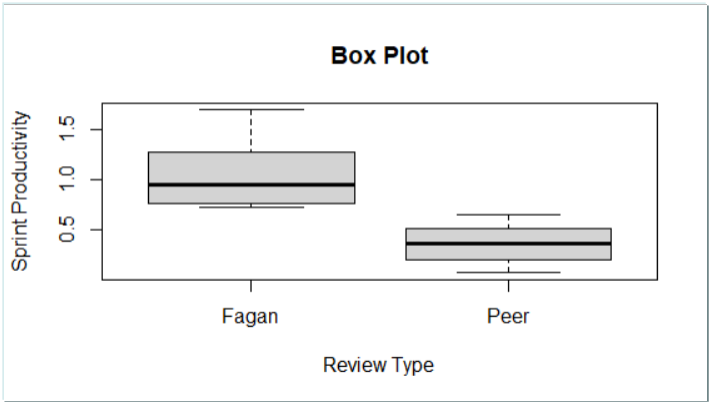
Preliminary Analysis: Developer Skill Vs Sprint Productivity



REGRESSION ANALYSIS

Regression with dummy variables

Preliminary Analysis: Review Type Vs Sprint Productivity

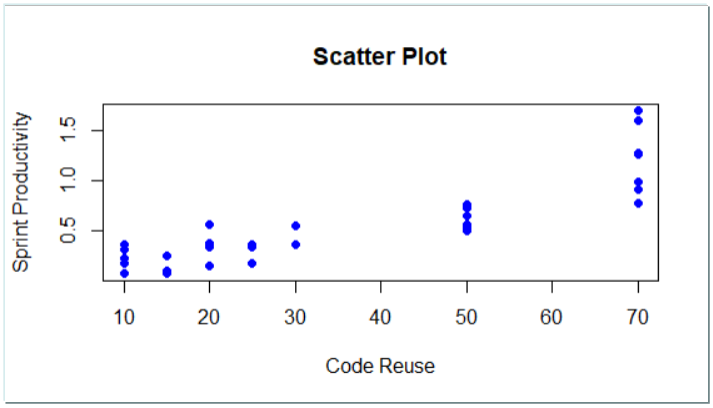


277

REGRESSION ANALYSIS

Regression with dummy variables

Preliminary Analysis: Review Type Vs Sprint Productivity



278

REGRESSION ANALYSIS

Regression with dummy variables – Output

```
mymodel = lm(SP ~ skill + review + reuse)
summary(mymodel)
```

Multiple R^2	0.9309
Adjusted R^2	0.9214
F Statistics	97.69
P value	0.00

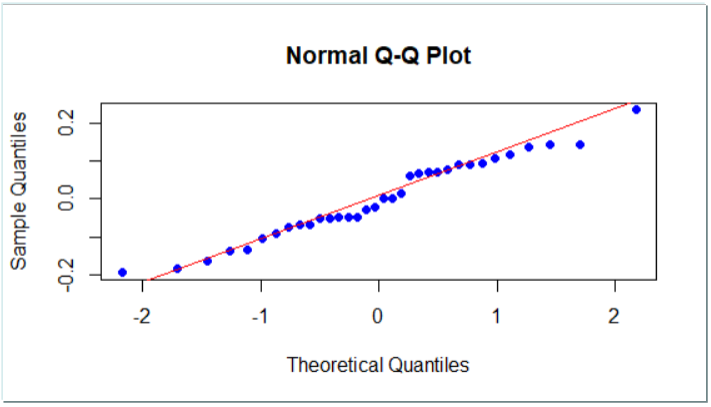
	Estimate	Std. Error	t value	p value
(Intercept)	0.405513	0.10296	3.939	0.00047
skill2-Experienced	0.198892	0.08048	2.471	0.01958
skill3-Master	0.799904	0.12452	6.424	0.00000
reviewPeer	-0.21398	0.07043	-3.038	0.0050

279

REGRESSION ANALYSIS

Regression with dummy variables – Model Adequacy Checks

Residual Analysis – Normal Q-Q Plot of Residuals



280

Indian Statistical Institute

REGRESSION ANALYSIS

Regression with dummy variables – Model Adequacy Checks

Residual Analysis – Normal Test of Residuals

Statistic	Value
w	0.97172
p_value	0.5106

Model accuracy Measures – MSE and RMSE

Statistic	Value
Mean Square Error (MSE)	0.011
Root Mean Square Error (RMSE)	0.1049

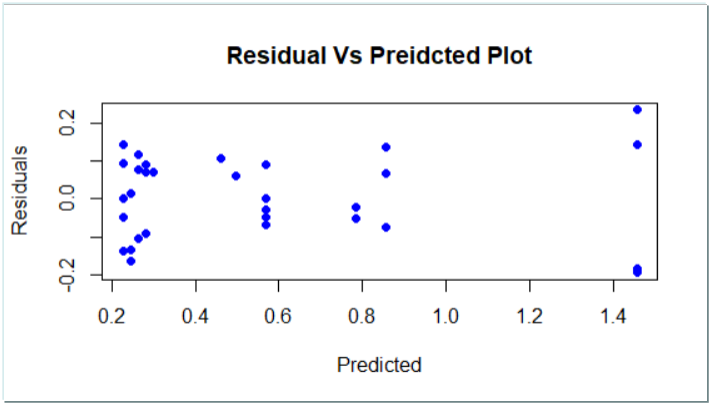
281

Indian Statistical Institute

REGRESSION ANALYSIS

Regression with dummy variables – Model Adequacy Checks

Residual Analysis – Residuals Vs Predicted Plot



282

REGRESSION ANALYSIS

Regression with dummy variables – Cross validation

```
library(boot)
attach(mydata)
set.seed(1)
```

LOOCV

```
mymodel = glm(Sprint_Productivty ~ factor(Developer_Skill) + factor(Review_Type)
+ Reuse)
mycv = cv.glm(mydata, mymodel)
round(mycv$delta[1],4)
```

K-fold Cross Validation

```
mymodel = glm(Sprint_Productivty ~ factor(Developer_Skill) + factor(Review_Type)
+ Reuse)
mycv = cv.glm(mydata, mymodel, K = 4)
round(mycv$delta[1],4)
```


REGRESSION ANALYSIS

Regression with dummy variables – Cross validation

Statistic	Training	LOOCV	K-fold CV
MSE	0.011	0.0167	0.0153
RMSE	0.1049	0.1292	0.1237

BINARY LOGISTIC REGRESSION

BINARY LOGISTIC REGRESSION

Used to develop models when the output or response variable y is binary

The output variable will be binary, coded as either success or failure

Models probability of success p which lies between 0 and 1

Linear model is not appropriate

$$p = \frac{e^{a+b_1x_1+b_2x_2+---+b_kx_k}}{1+e^{a+b_1x_1+b_2x_2+---+b_kx_k}}$$

p: probability of success
x_i's : independent variables
a, b_1, b_2, ---: coefficients to be estimated

If estimate of $p \geq 0.5$, then classified as success, otherwise as failure

BINARY LOGISTIC REGRESSION

Usage: When the dependant variable (*y* variable) is binary

Example: A study was performed to investigate new automobile purchases. A sample of 20 families was selected. Each family was surveyed to determine the age of their oldest vehicle and their total family income. A follow-up survey was conducted 6 months later to determine if they had actually purchased a new vehicle during that time period ($y = 1$ indicates yes and $y = 0$ indicates no). The data collected is given in Binary_Logistic file. Develop a model to predict whether a family would purchase a new automobile?

Indian Statistical Institute

BINARY LOGISTIC REGRESSION

Usage: When the dependant variable (*y* variable) is binary

Making the dataset as dataframe

```
mydata = as.data.frame(mydata)
```

Importing caret library

```
library(caret)
```

Separating x and y variables

```
x = mydata[,1:2]
y = mydata$New_Vehicle_Purchased
```

Converting y variable to factor

```
y = factor(y)
```

BINARY LOGISTIC REGRESSION

Usage: When the dependant variable (*y* variable) is binary

Density Plot

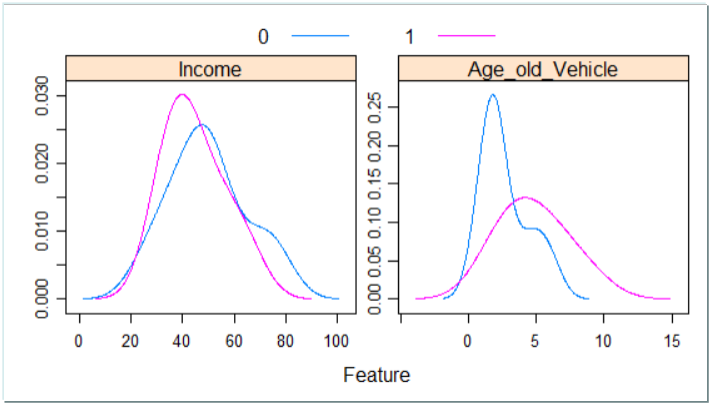
```
featurePlot(x, y, plot = "density", auto.key = list(columns = 2),
        scales = list(x = list(relation = "free"),
                      y = list(relation = "free")),
        adjust = 1.5, pch = "")
```

Box Plot

```
featurePlot(x, y, plot = "boxplot", auto.key = list(columns = 2),
        scales = list(x = list(relation = "free"),
                      y = list(relation = "free")))
```

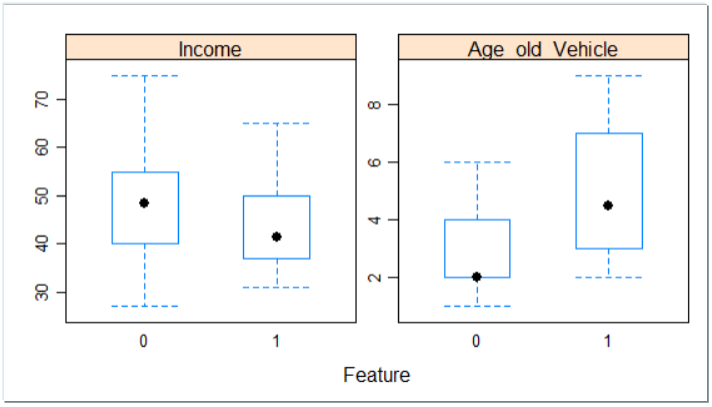

BINARY LOGISTIC REGRESSION

Density Plot



BINARY LOGISTIC REGRESSION

Box Plot



BINARY LOGISTIC REGRESSION

Usage: When the dependant variable (*y* variable) is binary

Reading the variables

```
> attach(mydata)
> income = Income
> age = Age_old_Vehicle
> purchase = New_Vehicle_Purchased
```

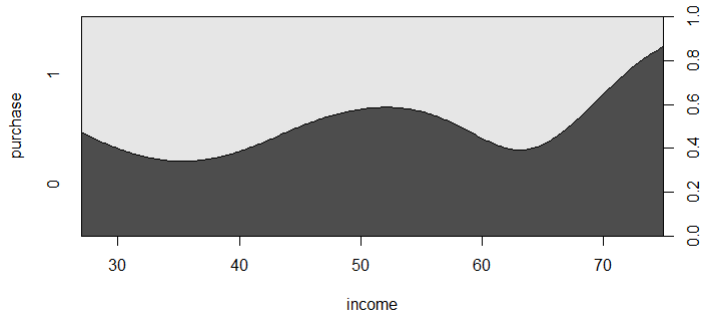
Converting response variable to discrete

```
> purchase = factor(purchase)
```

BINARY LOGISTIC REGRESSION

Checking relation between Xs and Y

> cdplot(purchase ~ income)

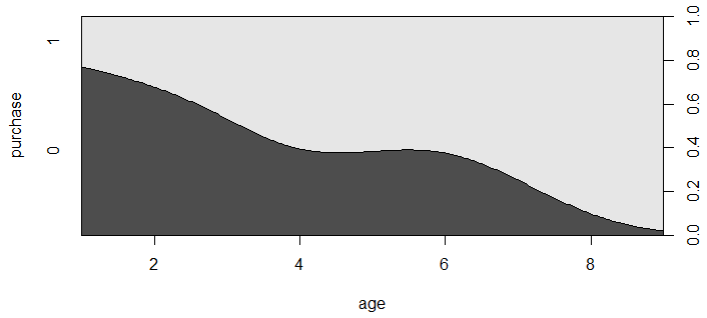


293

BINARY LOGISTIC REGRESSION

Checking relation between Xs and Y

> cdplot(purchase ~ age)



294

BINARY LOGISTIC REGRESSION

Perform Logistic regression

```
> mymodel = glm(purchase ~ income + age, family = binomial(logit))
> summary(mymodel)
```

	Estimate	Std. Error	z value	p-value
(Intercept)	-7.04706	4.67423	-1.508	0.132
Income	0.07382	0.06371	1.159	0.247
age	0.98789	0.52737	1.873	0.061

BINARY LOGISTIC REGRESSION

Perform Logistic regression - Anova

```
> anova(mymodel, test = "Chisq")
```

	Df	Deviance	Residual Df	Residual Deviance	p-value
NULL	19	27.726			
Income	1	0.7349	18	26.991	0.39129
Age	1	5.9094	17	21.081	0.01506

Since p-value < 0.05 for Age only
redo the modelling with significant factors only

BINARY LOGISTIC REGRESSION

Perform Logistic regression - Modified

	Estimate	Std. Error	z	p-value
(Intercept)	-2.0122	1.1142	-1.806	0.0709
Age	0.556	0.2878	1.932	0.0534

The model is

$$p(y) = \frac{e^{-2.0122+0.556\,Age}}{1+e^{-2.0122+0.556\,Age}}$$

BINARY LOGISTIC REGRESSION

Model significance test

```
> nullmodel = glm(purchase ~ 1, family = binomial(logit))
> anova(nullmodel, mymodel, test = "Chisq")
```

Model	Residual df	Residual deviance	df	Deviance	p-value
1	19	27.726	1	5.1625	0.02308
2	18	22.563			

Remark:
The p-value < 0.05, the model is significant at 5% level

BINARY LOGISTIC REGRESSION

Fitted Values and residuals

```
> pred = predict(mymodel, type = "response")
> predclass = ifelse(pred > 0.5, "1", "0")
```

New_Vehicle_Purchased	Predicted	Predclass
0	0.2890	0
0	0.5527	1
1	0.4148	0
1	0.2890	0
0	0.2890	0
1	0.6830	1
1	0.8676	1
1	0.2890	0
0	0.2890	0
0	0.1890	0
1	0.6830	1
1	0.8676	1
1	0.5527	1
0	0.2890	0
1	0.9522	1
0	0.2890	0
1	0.5527	1
0	0.1890	0
0	0.6830	1
0	0.7898	1

BINARY LOGISTIC REGRESSION

Model Evaluation

```
> mytable = table(New_Vehicle_Purchased, predclass)
> mytable
> prop.table(mytable)*100 predclass)
```

	Predicted Count		Total
Actual Count	0	1	
0	7	3	10
1	3	7	10
Total	20	20	20

	Predicted %		Total
Actual %	0	1	
0	35	15	50
1	15	35	50
Total	43	50	100

Statistics	Value
Accuracy %	70
Error %	30

Accuracy of ≥ 80 % is good

BINARY LOGISTIC REGRESSION

Cross Validation

```
attach(mydata)
mymodel = glm(factor(New_Vehicle_Purchased) ~ Age_old_Vehicle,
              family = binomial(logit))
```

LOOCV

```
mycv = cv.glm(mydata, mymodel)
mycv$delta[1]
```

K-fold Cross Validation

```
mycv = cv.glm(mydata, mymodel, K = 5)
mycv$delta[1]
```


BINARY LOGISTIC REGRESSION

Cross Validation

Statistic	Training	LOOCV	K-fold CV
Accuracy %	70	76.31	77.67
Misclassification %	30	23.69	22.33

BINARY LOGISTIC REGRESSION

Addition Performance Measures

	Predicted Count	
Actual Count	Negative (0)	Positive (1)
Negative (0)	True Negative (TN)	False Positive (FP)
Positive (1)	False Negative (FN)	True Positive (TP)

$$Sensitivity\ or\ Recall = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$Precision = \frac{TP}{TP + FP}$$

$$F - Measure = \frac{2\ x\ Precision\ x\ Recall}{Precision + Recall}$$

BINARY LOGISTIC REGRESSION

Addition Performance Measures

	Predicted Count	
Actual Count	0	1
0	7	3
1	3	7

Statistic	Value
Sensitivity	0.70
Specificity	0.70
Precision	0.70
F-Measure	0.70

CLASSIFICATION and REGRESSION TREE

CLASSIFICATION AND REGRESSION TREE

Objective

To develop a predictive model to classify dependant or response metric (y) in terms of independent or exploratory variablesxs).

When to Use

xs : Continuous or discrete

y : Discrete or continuous

CLASSIFICATION AND REGRESSION TREE

Classification Tree

When response y is discrete

Method = “class”

Regression Tree

When response y is numeric

Method = “anova”

CLASSIFICATION AND REGRESSION TREE

Example:

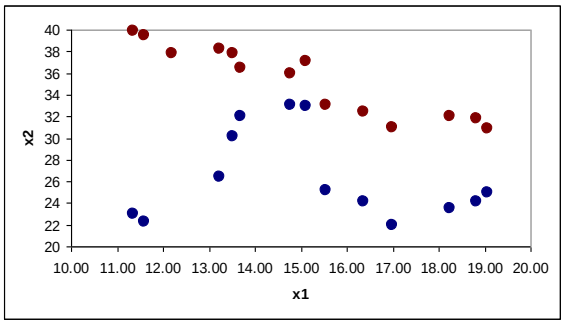
Attribute 1	x1
Attribute 2	x2
Label : y	y1 (Red) , y2 (Blue)

x1	x2	Y	x1	x2	Y
11.35	23	Blue	11.85	39.9	Red
11.59	22.3	Blue	12.09	39.5	Red
12.19	24.5	Blue	12.69	37.8	Red
13.23	26.4	Blue	13.73	38.2	Red
13.51	30.2	Blue	14.01	37.8	Red
13.68	32	Blue	14.18	36.5	Red
14.78	33.1	Blue	15.28	36	Red
15.11	33	Blue	15.61	37.1	Red
15.55	25.2	Blue	16.05	33.1	Red
16.37	24.1	Blue	16.87	32.4	Red
16.99	22	Blue	17.49	31	Red
18.23	23.5	Blue	18.73	32	Red
18.83	24.1	Blue	19.33	31.8	Red
19.06	25	Blue	19.56	30.9	Red

CLASSIFICATION AND REGRESSION TREE

Example:

Attribute 1	x1
Attribute 2	x2
Label : y	y1 (Red) , y2 (Blue)

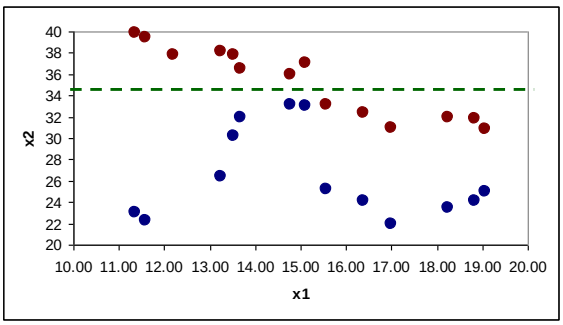


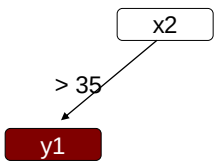
309

CLASSIFICATION AND REGRESSION TREE

Example:

Attribute 1	x1
Attribute 2	x2
Label : y	y1 (Red) , y2 (Blue)



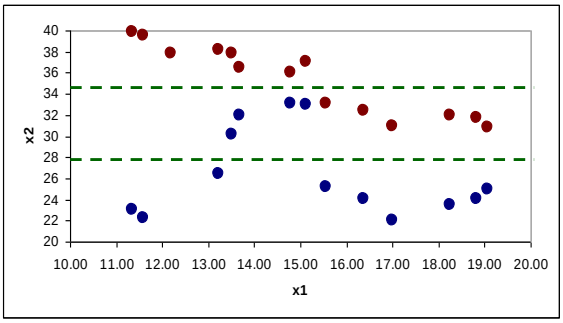


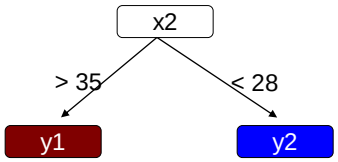
310

CLASSIFICATION AND REGRESSION TREE

Example:

Attribute 1	x1
Attribute 2	x2
Label : y	y1 (Red) , y2 (Blue)



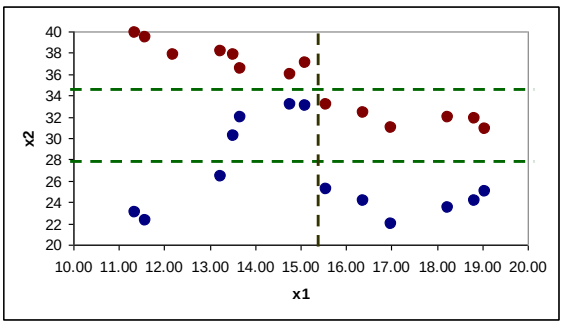


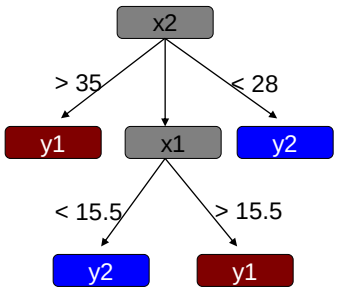
311

CLASSIFICATION AND REGRESSION TREE

Example:

Attribute 1	x1
Attribute 2	x2
Label : y	y1 (Red) , y2 (Blue)





312

CLASSIFICATION AND REGRESSION TREE

Example: Rules

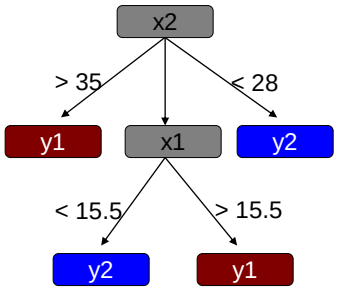
Attribute 1	x1
Attribute 2	x2
Label : y	y1 (Red) , y2 (Blue)

If x2 > 35 then y = y1

If x2 < 28, then y = y2

If 28 > x2 > 35 & x1 > 15.5, then y = y1

If 28 > x2 > 35 & x1 < 15.5, then y = y2



313

CLASSIFICATION AND REGRESSION TREE

Challenges

How to represent the entire information in the dataset using minimum number of rules?

How to develop the smallest tree?

Solution

Select the variable with maximum information (highest relation with y) for first split

314

CLASSIFICATION AND REGRESSION TREE

Example: A marketing company wants to optimize their mailing campaign by sending the brochure mail only to those customers who responded to previous mail campaigns. The profile of customers are given below. Can you develop a rule to identify the profile of customers who are likely to respond (Mail_Respond.csv)?

CLASSIFICATION AND REGRESSION TREE

Example: A marketing company wants to optimize their mailing campaign by sending the brochure mail only to those customers who responded to previous mail campaigns. The profile of customers are given in mail_respond.csv? Can you develop a rule to identify the profile of customers who are likely to respond?

Number of variables = 4

SL No	Variable Name	Number of values
1	District	3
2	House Type	3
3	Income	2
4	Previous Customer	2

Total Combination of Customer Profiles = 3 x 3 x 2 x 2 = 36

CLASSIFICATION AND REGRESSION TREE

Read file and variables

```
> mydata = Mail_Respond
> house = mydata$House_Type
> district = mydata$District
>  income = mydata$Income
> prev = mydata$Previous_Customer
> outcome = mydata$Outcome
```


CLASSIFICATION AND REGRESSION TREE

Develop the model

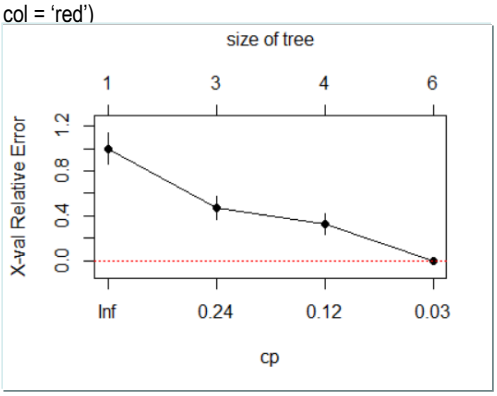
```
> library(rpart)
> library(rpart.plot)
> mymodel = rpart(outcome ~ district + house + income + prev, method = 'class', control = rpart.control(minsplit = 2))
```

Note: When response is categorical, method = “class”, when response is numeric, methos = “anova”

CLASSIFICATION AND REGRESSION TREE

Cross validation and identification of cp

> plotcp(mymodel, pch = 19, col = 'red')



Optimum cp = 0.03
Corresponding to minimum cross validation relative error

319

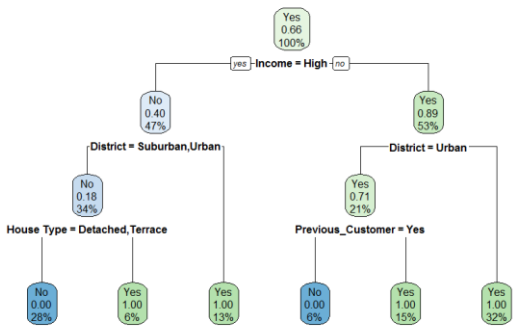
CLASSIFICATION AND REGRESSION TREE

Pruning the model

> mymodel = prune(mymodel, cp = 0.03)

Plot the model

> rpart.plot(mymodel)



320

CLASSIFICATION AND REGRESSION TREE

Print the model

> mymodel

Income	District	House	Previous	Outcome
High	Suburban, Urban	Detached,Terrace		No
		Semi-detached		Yes
	Rural			Yes
Low	Urban		Yes	No
			No	Yes
	Rural, Suburban			Yes

CLASSIFICATION AND REGRESSION TREE

Model Accuracy measures

> pred = predict(mymodel, type = "class")

> mytable =table(outcome, pred)

> prop.table(mytable)*100

Actual Vs predicted: %

Actual	Predicted	
	No	Yes
No	34	0
Yes	0	66

Accuracy = 34 + 66 = 100%

CLASSIFICATION AND REGRESSION TREE

Exercise 1: Develop a tree based model for predicting whether the customer will take pep using the customer profile data given in bank-data.csv? Use 80% of data to develop the model and validate the model using the remaining 20% of data?

CLASSIFICATION AND REGRESSION TREE

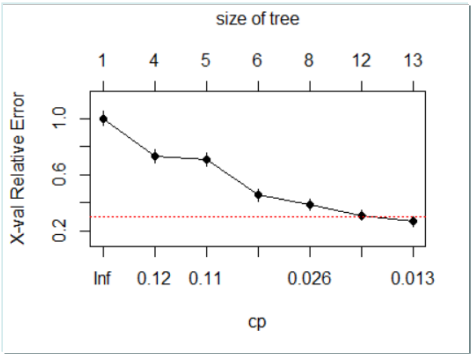
Exercise 1: Develop a tree based model for predicting whether the customer will take pep using the customer profile data given in bank-data.csv? Use 80% of data to develop the model and validate the model using the remaining 20% of data?

```
> set.seed(1)
> sample = sample(2, nrow(mydata), replace =TRUE, prob = c(0.8, 0.2))
> training = mydata[sample == 1,]
> test = mydata[sample == 2, ]
```

```
> attach(training)
> library(rpart)
> library(rpart.plot)
> mymodel = rpart(pep ~ age + sex + region + income + married + children + car + save_act +
current_act + mortgage, method = "class", control = rpart.control(minsplit = 50), data = training)
> plotcp(mymodel)
```

CLASSIFICATION AND REGRESSION TREE

Exercise 1: Develop a tree based model for predicting whether the customer will take pep using the customer profile data given in bank-data.csv? Use 80% of data to develop the model and validate the model using the remaining 20% of data?



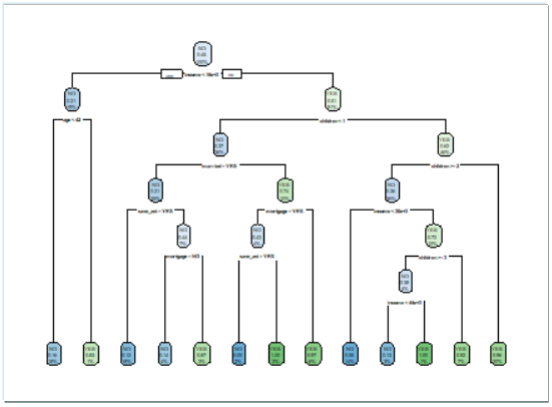
Optimum cp with minimum cross validation relative error = 0.013

```
> mymodel = prune(mymodel, cp = 0.013)
```


CLASSIFICATION AND REGRESSION TREE

Exercise 1: Develop a tree based model for predicting whether the customer will take pep using the customer profile data given in bank-data.csv? Use 80% of data to develop the model and validate the model using the remaining 20% of data?

```
rpart.plot(mymodel)
```



CLASSIFICATION AND REGRESSION TREE

Exercise 1: Develop a tree based model for predicting whether the customer will take pep using the customer profile data given in bank-data.csv? Use 80% of data to develop the model and validate the model using the remaining 20% of data?

```
> pred = predict(mymodel, type = "class", data = training)
> myable = table(training$pep, pred)
 > prop.table(mytable)*100
```

Actual Vs predicted: %

Actual	Predicted	
	No	Yes
No	52.27	2.27
Yes	7.44	38.02

Accuracy = 52.27 + 38.02 = 90.29 %

327

CLASSIFICATION AND REGRESSION TREE

Exercise 1: Develop a tree based model for predicting whether the customer will take pep using the customer profile data given in bank-data.csv? Use 80% of data to develop the model and validate the model using the remaining 20% of data?

```
> predtest = predict(mymodel, type = "class", newdata = test)
> mytesttable = table(test$pep, predtest)
> prop.table(mytesttable)*100
```

Actual Vs predicted: %

Actual	Predicted	
	No	Yes
No	49.14	4.31
Yes	5.17	41.38

Accuracy = 49.149 + 41.384 = 90.52 %

328

CLASSIFICATION AND REGRESSION TREE

Exercise 1: Develop a tree based model for predicting whether the customer will take pep using the customer profile data given in bank-data.csv? Use 80% of data to develop the model and validate the model using the remaining 20% of data?

Exporting and saving the model

```
> saveRDS(mymodel, file = "E:/ISI/BA-07/Course_Material/CARTmodel.rds")
```

Importing the model and prediction

```
> mynewmodel = readRDS("E:/ISI/BA-07/Course_Material/CARTmodel.rds")
> rpart.plot(mynewmodel)
> pred = predict(mynewmodel, newdata = test, type = "class")
> predprob = predict(mynewmodel, newdata = test, type = "prob")
> result = cbind(test, predprob, pred)
```

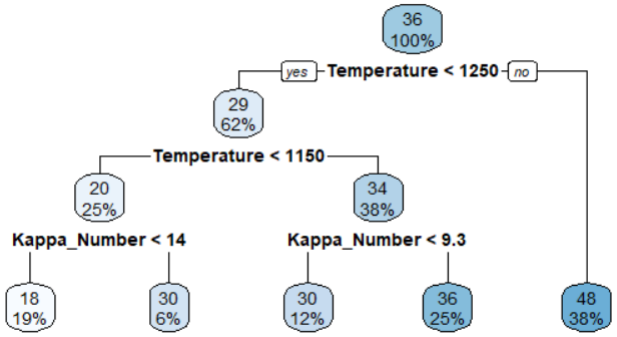

CLASSIFICATION AND REGRESSION TREE

Exercise 2: Develop a tree based model for predicting conversion using temperature and kappa number as factors. The data is given in Mult_Reg_Conversion.csv?

SL No.	Temperature	Kappa_Number	Conversion
1	1300	7.5	49
2	1300	9	50.2
3	1300	11	50.5
4	1300	13.5	48.5
5	1300	17	47.5
6	1300	23	44.5
7	1200	5.3	28
8	1200	7.5	31.5
9	1200	11	34.5
10	1200	13.5	35
11	1200	17	38
12	1100	5.3	15
12	1200	23	38.5
14	1100	7.5	17
15	1100	11	20.5
16	1100	17	29.5

CLASSIFICATION AND REGRESSION TREE

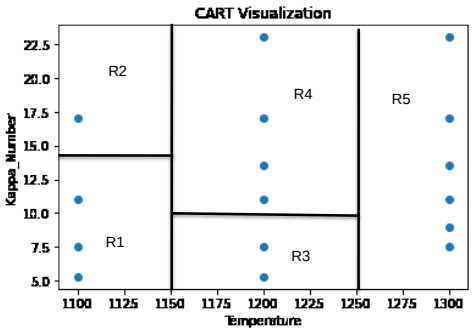
Exercise 2: Develop a tree based model for predicting conversion using temperature, time and kappa number as factors. The data is given in Mult_Reg_Conversion.csv?



331

CLASSIFICATION AND REGRESSION TREE

Exercise 2: Develop a tree based model for predicting conversion using temperature, time and kappa number as factors. The data is given in Mult_Reg_Conversion.csv?



Value estimation problem (y : Numeric)

Predicted value ($\hat{y}_{R_j}$) in a region R_j: Average of all values ($y_j \in R_j$) in the region

332

CLASSIFICATION AND REGRESSION TREE

Exercise 2: Develop a tree based model for predicting conversion using temperature, time and kappa number as factors. The data is given in Mult_Reg_Conversion.csv?

SL No.	Temperature	Kappa_Number	Conversion	Region	Predicted
12	1100	5.3	15	R1	17.5
14	1100	7.5	17		
15	1100	11	20.5		
16	1100	17	29.5	R2	29.5
7	1200	5.3	28	R3	29.75
8	1200	7.5	31.5		
9	1200	11	34.5	R4	36.5
10	1200	13.5	35		
11	1200	17	38		
12	1200	23	38.5		
1	1300	7.5	49	R5	48.37
2	1300	9	50.2		
3	1300	11	50.5		
4	1300	13.5	48.5		
5	1300	17	47.5		
6	1300	23	44.5		

333

Indian Statistical Institute

CLASSIFICATION AND REGRESSION TREE

Exercise 2: Develop a tree based model for predicting conversion using temperature, time and kappa number as factors. The data is given in Mult_Reg_Conversion.csv?

Procedure: Identification of splitting variable and value
Top down greedy approach
Binary recursive splitting
By minimizing residual sum of squares (Res_SS)

$$Res\_SS = \sum_{j=1}^{j} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2$$

334

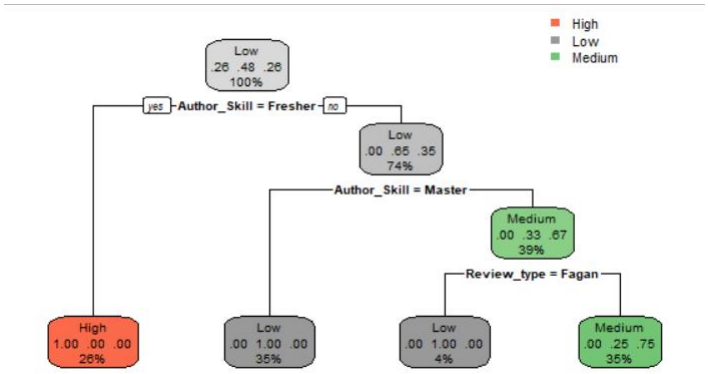
CLASSIFICATION AND REGRESSION TREE

Exercise 3: Develop a tree based model for predicting defect density using author skill and review type?

Author_Skill	Review_type	Defect_density
Fresher	Peer	High
Fresher	Peer	High
Fresher	Peer	High
Fresher	Peer	High
Fresher	Peer	High
Fresher	Peer	High
Experienced	Peer	Medium
Experienced	Peer	Medium
Experienced	Peer	Medium
Experienced	Peer	Medium
Experienced	Peer	Medium
Experienced	Peer	Medium
Experienced	Peer	Low
Experienced	Peer	Low
Experienced	Fagan	Low
Master	Peer	Low
Master	Peer	Low
Master	Fagan	Low
Master	Peer	Low
Master	Fagan	Low
Master	Peer	Low
Master	Fagan	Low
Master	Fagan	Low

CLASSIFICATION AND REGRESSION TREE

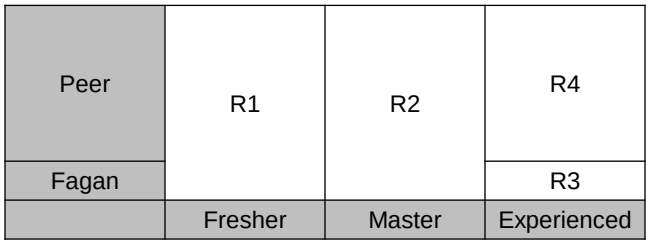
Exercise 3: Develop a tree based model for predicting defect density using author skill and review type?



Classification Problem:
Predicted class k is the one with m

CLASSIFICATION AND REGRESSION TREE

Exercise 3: Develop a tree based model for predicting defect density using author skill and review type?



Classification Problem:
Predicted class k in a region m: class with maximum number of response

$$\hat{k}_m = max_k(p_{mk})$$

p_{mk} : proportion of data points in m^{th} region that are from k^{th} class

337

CLASSIFICATION AND REGRESSION TREE

Exercise 3: Develop a tree based model for predicting defect density using author skill and review type?

Author_Skill	Review_type	Defect_density	Region	Response	Proportion	Predicted class
Fresher	Peer	High	R1	Low = 0 Medium = 0 High = 6	Low = 0 Medium = 0 High = 1	High
Fresher	Peer	High				
Fresher	Peer	High				
Fresher	Peer	High				
Fresher	Peer	High				
Fresher	Peer	High				
Master	Peer	Low	R2	Low = 8 Medium = 0 High = 0	Low = 1 Medium = 0 High = 0	Low
Master	Peer	Low				
Master	Fagan	Low				
Master	Peer	Low				
Master	Fagan	Low				
Master	Peer	Low				
Master	Fagan	Low				
Master	Fagan	Low				
Experienced	Fagan	Low	R3	Low = 1 Medium = 0 High = 0	Low = 1 Medium = 0 High = 0	Low
Experienced	Peer	Medium	R4	Low = 2 Medium = 6 High = 0	Low = 0.25 Medium = 0.75 High = 0	Medium
Experienced	Peer	Medium				
Experienced	Peer	Medium				
Experienced	Peer	Medium				
Experienced	Peer	Medium				
Experienced	Peer	Medium				
Experienced	Peer	Low				
Experienced	Peer	Low				

338

CLASSIFICATION AND REGRESSION TREE

Exercise 3: Develop a tree based model for predicting defect density using author skill and review type?

Procedure: Identification of splitting variable and value
Binary recursive splitting
By minimizing Gini index (G) or cross-entropy (D)

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

$$D = -\sum_{k=1}^{K} \hat{p}_{mk} log \hat{p}_{mk}$$

p_{mk} : proportion of data points in m^{th} region that are from k^{th} class

339

RANDOM FOREST
and
BAGGING

340

RANDOM FOREST

Improves predictive accuracy

Generates large number of bootstrapped trees

Classifies a new case using each tree in the new forest of trees

Final predicted outcome by combining the results across all of the trees

Regression tree – average

Classification tree – majority vote

BAGGING

- Uses trees as building blocks to construct more powerful prediction models
- Decision trees suffer from high variance

  If we split the data into two parts and construct two different trees for each half of the data, the trees can be quite different
- In contrast, a proceedure with low varaince will yield similar results if applied repeatedly to distinct datasets
- Bagging is a general purpose procedure for reducing the variance of a statistical learning method

BAGGING

Procedure

- Take many training sets from the population
- Build seperate prediction models using each training set
- Average the resulting predictions
- Averaging of a set of observatins reduce variance
- Different training datasets are taken using bootstrap sampling
- Generally bootstraped sample consists of two third of the observations and the model is tested on the remainng one third ofthe out of the bag observations

For discrete response – will take the majority vote instead of average

Major difference between bagging and Random Forest

Bagging generally uses all the p predictors while random forest uses $\sqrt{p}$ predictors

343

BAGGING

Example

Develop a model to predict the medain value of owner occupied homes using Bosten housing data ? Use 80% of the data to develop the model and validate the model using remaining 20% of the data?

R Code

```
> library(randomForest)
> set.seed(1)

> sampleid = sample(2, 506, replace = TRUE, prob = c(0.8, 0.2))
> train = mydata[sampleid == 1,]
> test = mydata[sampleid == 2,]
> attach(train)

> mymodel = randomForest(MEDV ~., data = train, mtry = 13, importance =
TRUE)
> mymodel
```

344

BAGGING

Example

Develop a model to predict the medain value of owner occupied homes using Bosten housing data ? Use 80% of the data to develop the model and validate the model using remaining 20% of the data?

Output

Statistic	Value
R^2	85.46
MSE	12.41
RMSE	3.5224

BAGGING

Example

Develop a model to predict the medain value of owner occupied homes using Bosten housing data ? Use 80% of the data to develop the model and validate the model using remaining 20% of the data?

R Code : Variable Importance

```
> importance(mymodel)
> varImpPlot(mymodel)
```

BAGGING

Example

Develop a model to predict the medain value of owner occupied homes using Bosten housing data ? Use 80% of the data to develop the model and validate the model using remaining 20% of the data?

Output : Variable Importance

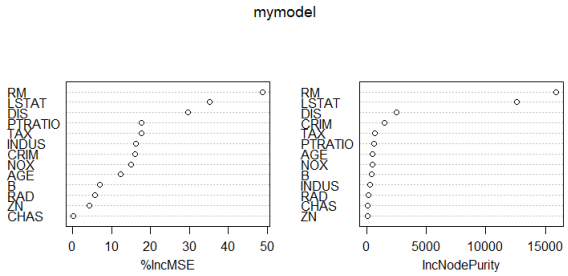
Variable	%IncMSE	IncNodePurity
CRIM	15.95744	1495.52571
ZN	4.302619	35.76408
INDUS	16.25835	281.13619
CHAS	0.194061	45.12435
NOX	14.9107	473.83672
RM	48.75136	15900.57962
AGE	12.31431	482.45746
DIS	29.67157	2487.69899
RAD	5.706936	116.66253
TAX	17.55344	670.70542
PTRATIO	17.62702	583.70698
B	7.019372	406.87729
LSTAT	35.18478	12577.58833

BAGGING

Example

Develop a model to predict the medain value of owner occupied homes using Bosten housing data ? Use 80% of the data to develop the model and validate the model using remaining 20% of the data?

Output : Variable Importance



Note: RM and LSTAT are the most important variables

BAGGING

Variable Importance

Regression Tree

The total amount of residual sum of squares is decreased due to splits over a given explanatory variable, averaged over all the trees.

Classification Tree

The total amount of Gini index is decreased due to splits over a given explanatory variable, averaged over all the trees.

A large value indicates an important explanatory variable

349

BAGGING

Example

Develop a model to predict the medain value of owner occupied homes using Bosten housing data ? Use 80% of the data to develop the model and validate the model using remaining 20% of the data?

R Code: Validation

```
> predtest = predict(mymodel, newdata = test)
> restest = test$MEDV - predtest
> mse = mean(restest^2)
> mse
> rmse = sqrt(mse)
> rmse
```

350

BAGGING

Example

Develop a model to predict the medain value of owner occupied homes using Bosten housing data ? Use 80% of the data to develop the model and validate the model using remaining 20% of the data?

Output : Validation

Statistic	Value
MSE	10.0247
RMSE	3.1662

Random Forest Model: Don't specify the number of explanatory variables
mymodel = randomForest(MEDV ~., data = train, importance = TRUE)

BAGGING

Exercise 1

Develop a model to predict whether a customer will take personal equity plan or not using bank-data .csv. Use 80% of the data to develop the model and validate the model using remaining 20% of the data?

Note: For classification, discretize the response variable y using factor() statement

Indian Statistical Institute

Faculty Development Program

on

Data Science

Thank You

353