Training Program on

Statistical Techniques

for

Data Mining & Business Analytics

June 5-7 and 20-21 2015

Indian Statistical Institute (ISI)

- The ***Indian Statistical Institute*** is a non-profit distributing scientific organization registered under the Societies Registration Act.
- It is declared by an act of parliament as an Institute of National Importance.
- Over the years the Institute has grown as a multi-disciplinary organization.
- It functions as a University in educational programmes and degree awarding activities; as a corporation in undertaking large scale projects; as a firm of consultants to industries to improve Quality, Reliability and Efficiency and as a meeting place of Scientists, Economists and Literary figures from all parts of the world.

Role & Function of SQC & OR DIVISION

- **The pioneer and leader in blending statistical theory with practice and institutionalizing the continuous improvement process into a sustaining system.**
- **To strengthen national economy through continual search for excellence in Quality.**
- **To play a leading role in dissemination of new concepts, methods and techniques in the improvement of Quality and Productivity.**
- **To develop highly skilled professionals who are capable of self actualization..**
- **To help industries in their efforts to cope up with the growing challenge of global competition through implementation of quality management system.**
- **To continually develop and improve methodologies through applied research efforts to attain International Standards in services provided.**

Programme Objectives

- **Describe a practical approach for making sense out of data**

- **To understand –**
  a. **How to summarize and interpret the data,**
  b. **How to identify patterns, relationships in the data,**
  c. **How to make predictions from the data and**
  d. **How to avoid common pitfall.**

CONTENTS

Introduction

Introduction

Some Issues:-

- **Predicting the buying behavior of your prospects.**
- **Identifying first-mover advantage by introducing new products and services.**
- **Evaluating the impact of marketing campaigns/advertisements.**
- **Understanding the trend and reason of customer/ employee attrition.**
- **Predict likely failures of critical equipment and processes.**
- **Correlating process input with output.**

Some Issues:-

- Predicting the buying behavior of your prospects.
- Identifying first-mover advantage by introducing new products and services.
- Evaluating the impact of marketing campaigns/advertisements.
- Understanding the trend and reason of customer/ employee attrition.
- Predict likely failures of critical equipment and processes.
- Correlating process input with output.

Business/ Data Analytics:-

- **The data derive meaningful trends or intriguing findings that were not previously seen or empirically validated**
- **Data analytics enables quick decisions or help change policies due to trends observed**
- **Accumulation of raw data captured from various sources (i.e. discussion boards, emails, exam logs, chat logs in e-learning systems) can be used to identify fruitful patterns and relationships (Bose, 2009)**
- **Exploratory visualization – uses exploratory data analytics by capturing relationships that are perhaps unknown or at least less formally formulated**
- **Confirmatory visualization - theory-driven**

Data Analytics vs. Statistical Analysis

Data Analytics

- Utilizes data mining techniques
- Identifies inexplicable or novel relationships/trends
- Seeks to visualize the data to allow the observation of relationships/trends

Statistical Analysis

- Utilizes statistical and/or mathematical techniques
- Used based on theoretical foundation
- Seeks to identify a significant level to address hypotheses or Research Questions

What is Business Analytics

Business analytics (BA) is

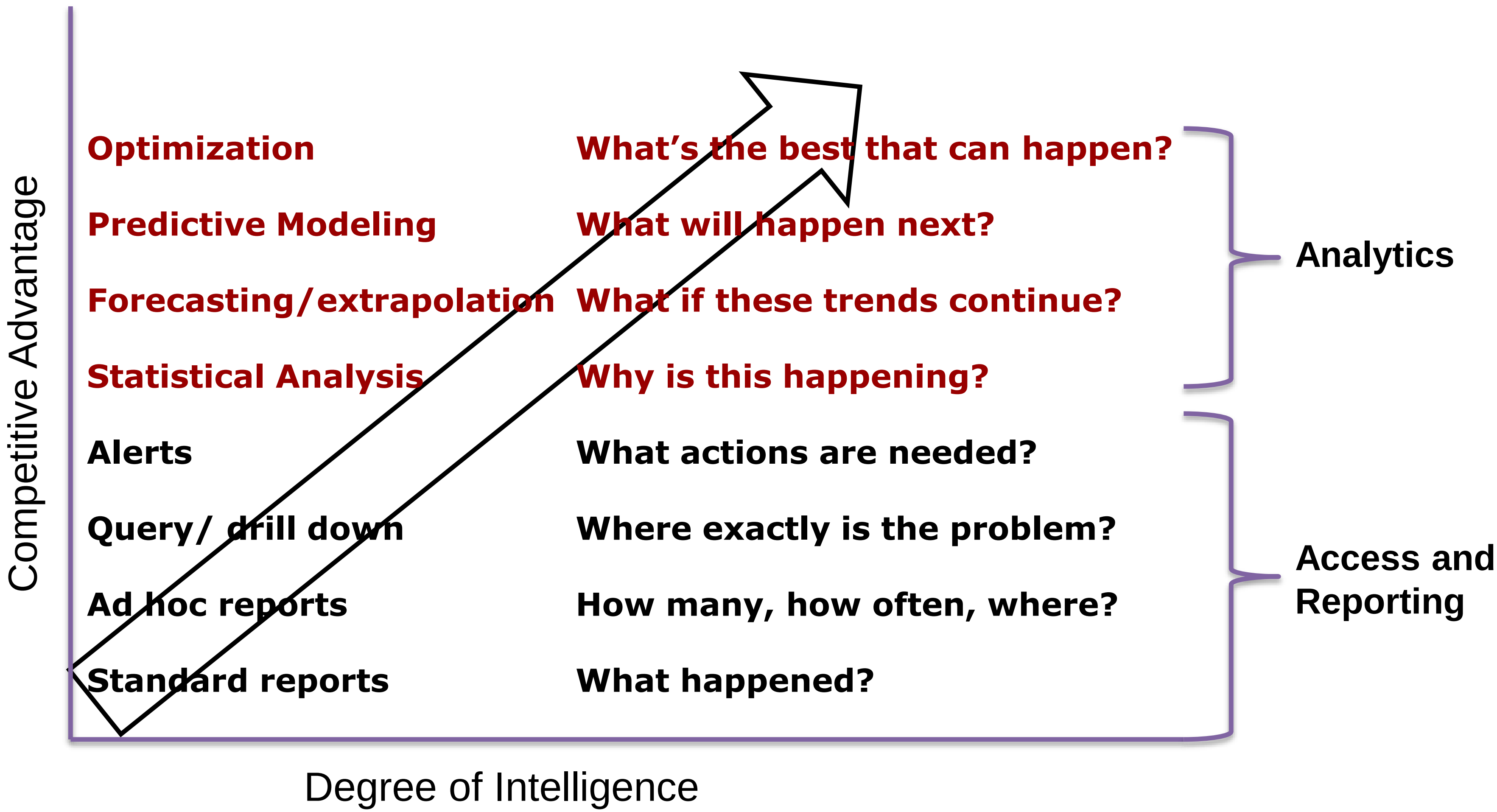
- **Translating data into information to make informed decisions.**

- **The practice of iterative, methodical exploration of an organization's data with emphasis on statistical analysis for data-driven decision making**

- **The discovery and communication of meaningful patterns in data using tabulation and visualization techniques to communicate insights. It relies on the simultaneous application of computer programming and quantitative techniques to quantify performance.**

- **The extensive use of data, statistical and quantitative analysis, explanatory and predictive models, and fact-based management to drive decisions and actions.**

Business intelligence and analytics



Source: Davenport, Thomas H.; Harris, Jeanne G. Competing on Analytics: The New Science of Winning Harvard Business Press.

FUNDAMENTALS

of

STATISTICS

- **What is Meant by Statistics?**

- ***Statistics*** **is the science of collecting, organizing, presenting, analyzing, and interpreting numerical data for the purpose of assisting in making a more effective decision.**

- **Who Uses Statistics?**

- **Statistical techniques are used extensively by marketing, accounting, quality control, consumers, professional sports people, hospital administrators, educators, politicians, physicians, etc...**

Types of Statistics

- **Descriptive Statistics:** Methods of organizing, summarizing, and presenting data in an informative way.

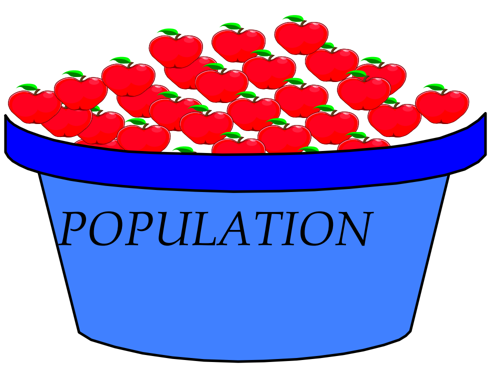
EXAMPLE : According to Consumer Reports, Whirlpool washing machine owners reported 9 problems per 100 machines during 2007. The statistic 9 describes the number of problems out of every 100 machines.

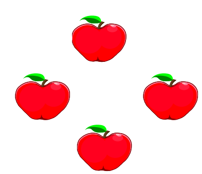
- **Inferential Statistics:** A decision, estimate, prediction, or generalization about a population, based on a sample
  - A **population** is a collection of all possible individuals, objects, or measurements of interest.
  - A **sample** is a portion, or part, of the population of interest.

EXAMPLE : TRP - As per research organization the programme, ".." has the highest viewer base.

Population and Sample

- The entire set of items is called the *Population.*
- The small number of items taken from the population to make a judgment of the population is called a *Sample*.
- The numbers of samples taken to make this judgment is called *Sample size.*





Sample of Size Four

Sampling must be representative to enable solid conclusions.

Types of Variables (Data)

Qualitative or Attribute variable: The characteristic or variable being studied is nonnumeric.

- **EXAMPLES: Gender, religious affiliation, type of automobile owned, state of birth, eye color.**

Quantitative variable: the variable can be reported numerically.

- **EXAMPLE: balance in your checking account, minutes remaining in class, number of children in a family.**

- **Quantitative variables can be classified as either discrete or continuous...**
- **Discrete variables: can only assume certain values and there are usually "gaps" between values. Sometimes it is know as Attributes**

- Data generated by
  - Counting or classifying the items into different groups based on some criteria
  - No physical measurement is involved
  - Not measured on a continuous scale
  - Nominal/ Ordinal / Binary

Examples:

Gender, Shade variation, Surface defects etc.

On Time Delivery of Tasks, Defect free Delivery of Tasks, Defects injected, Defects detected etc.

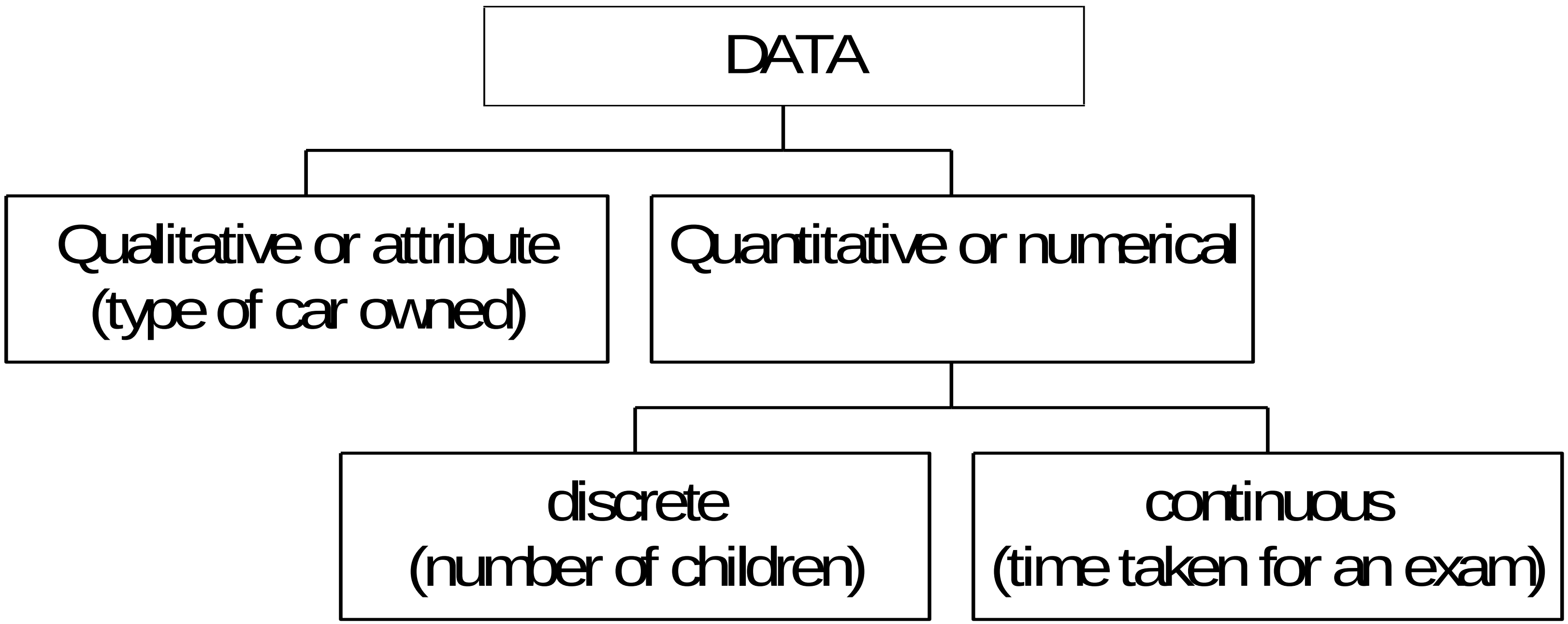
FUNDAMENTALS OF STATISTICS

- **Continuous variables: can assume any value within a specific range.**

- Data generated by
  - Physically measuring the characteristic
  - Generally using an instrument
  - Assigning an unique value to each item measured
  - Measurable
  - Expressed on continuous scale of measurement

- Example
  - Hardness, Strength, Weight, Diameter, Cycle Time etc

Summary of Types of Variables (Data)



STRIVE TO COLLECT QUANTITATIVE DATA

Exercise : Which of the Below are Continuous and Discrete Data?

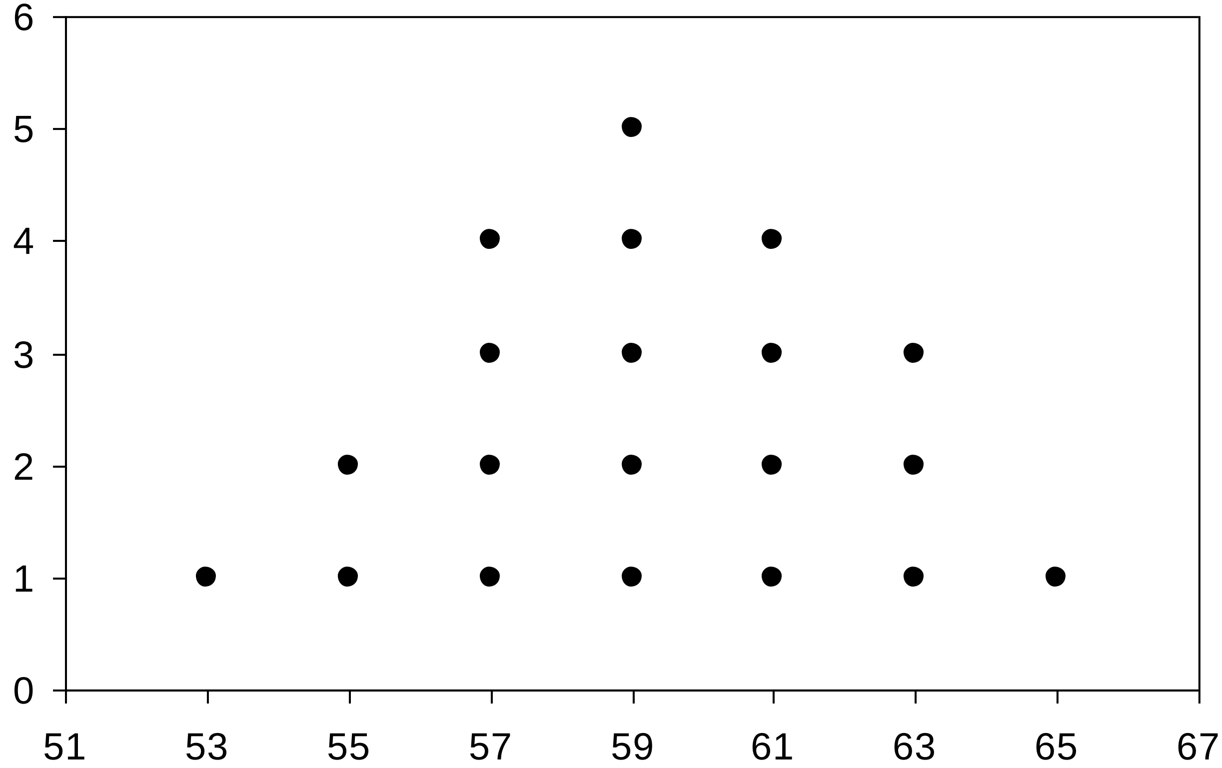
1. Time taken to process a purchase order
2. Units sold in a week
3. TAT (Cycle or Lead) for issuing invoice
4. Number of protocol violation during call
5. Document scrutinized during an hour
6. Number of printing defects on a shipping label
7. Number of typos per Sales Contract
8. Average response time to customer special orders
9. Account Receivable
10. Amount of time to close an account
11. Number of new hires per 100 applicants
12. Productivity of Agent

Description of sample data

The monthly credit card expenses of an individual in 1000 rupees is given below. Kindly summarize the data

Month	Credit Card Expenses	Month	Credit Card Expenses
1	55	11	63
2	65	12	55
3	59	13	61
4	59	14	61
5	57	15	57
6	61	16	59
7	53	17	61
8	63	18	57
9	59	19	59
10	57	20	63

Summarization of sample data



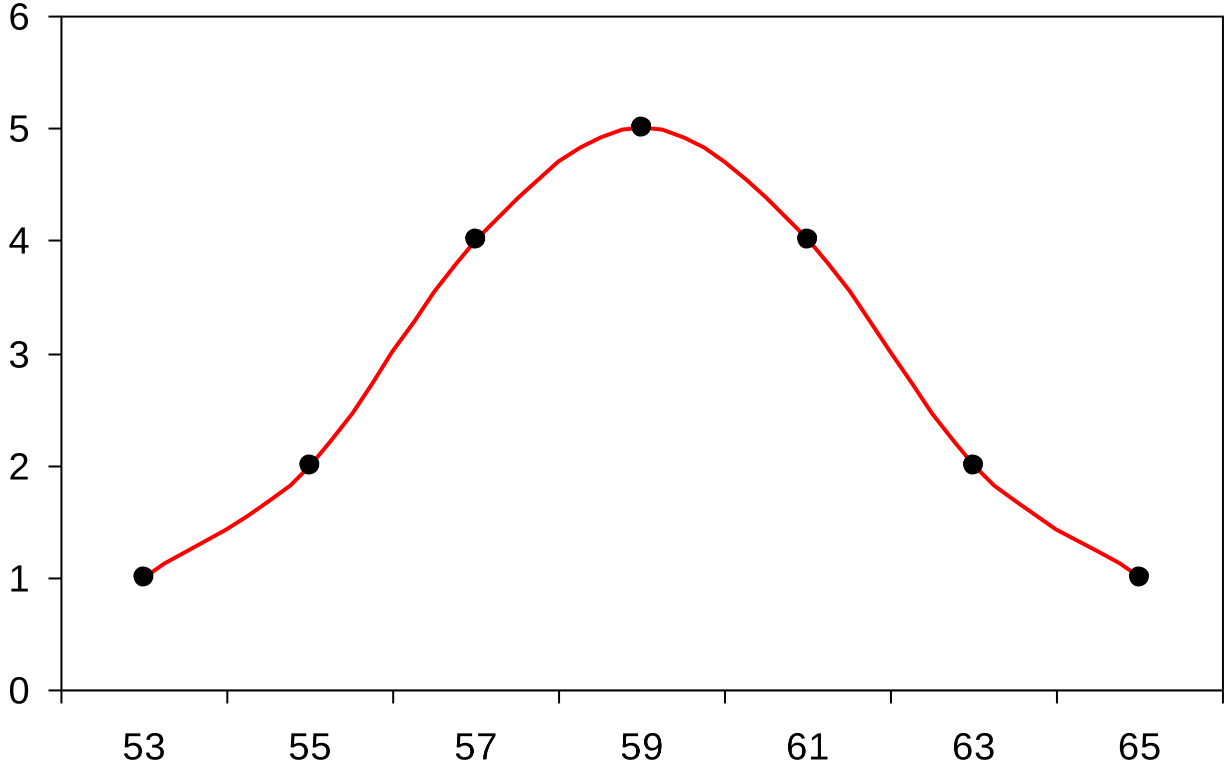
Summary: 1. Central tendency

2. Dispersion or variation

3. Shape or distribution

FUNDAMENTALS OF STATISTICS

Summarization of sample data



Summary: 1. Central tendency

2. Spread or variation

3. Shape or distribution

FUNDAMENTALS OF STATISTICS

Variable Data: Measure of Central tendency

Sample Average:

- Numerical value indicating the centre of data set
- Sum of all data points / Total number of data points

Suppose x_1, x_2, - - - x_n be the data set, then

$$\text{Sample Average} = \overline{X} = \frac{X_1 + X_2 + \cdots\cdots + X_n}{n} = \sum_{i=1}^{n} \frac{x_i}{n}$$

Summarization of sample data: Credit Card Expenses

Sample Average: :Sum of all data points / Total number of data points

= (55 + 65 + 59 + 59 + 57 + 61 + 53 + 63 + 59 + 57 + 63 + 55 + 61 + 61 + 57 + 59 + 61 + 57 + 59 + 63) / 20

= 1184 / 20 = 59.2

Interpretation

On an average, the individual spends Rs. 59200 through credit card monthly

Summarization of sample data: Measure of Central tendency

Sample Median:

Value which divides the data set arranged in ascending or descending order of values into two equal halves

Case 1: Total number of values in data set is odd

Median: Middle Value

Case 2: Total number of values in data set is even

Median: Average of two middle values

Credit Card Expenses

Median = ?

Summarization of sample data: Measure of Central tendency

Sample Median: Credit Card Expenses

Month	Credit Card Expenses	Month	Credit Card Expenses
1	53	11	59
2	55	12	59
3	55	13	61
4	57	14	61
5	57	15	61
6	57	16	61
7	57	17	63
8	59	18	63
9	59	19	63
10	59	20	65

Median = 59

Interpretation

50% of the months the credit card expenses are less than or equal to Rs. 59,000/-

Summarization of sample data : Measure of Central tendency

Sample Mode:

- The value which occurs maximum number of times in the data set

Example: Credit Card Expenses

Values	No. of Occurrences
53	1
55	2
57	4
59	5
61	4
63	3
65	1
Total	20

Mode = 59

Interpretation

Maximum number of months, the credit card expenses is equal to Rs. 59,000/-

Summarization of sample data : Measure of Variation or dispersion

Sample 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

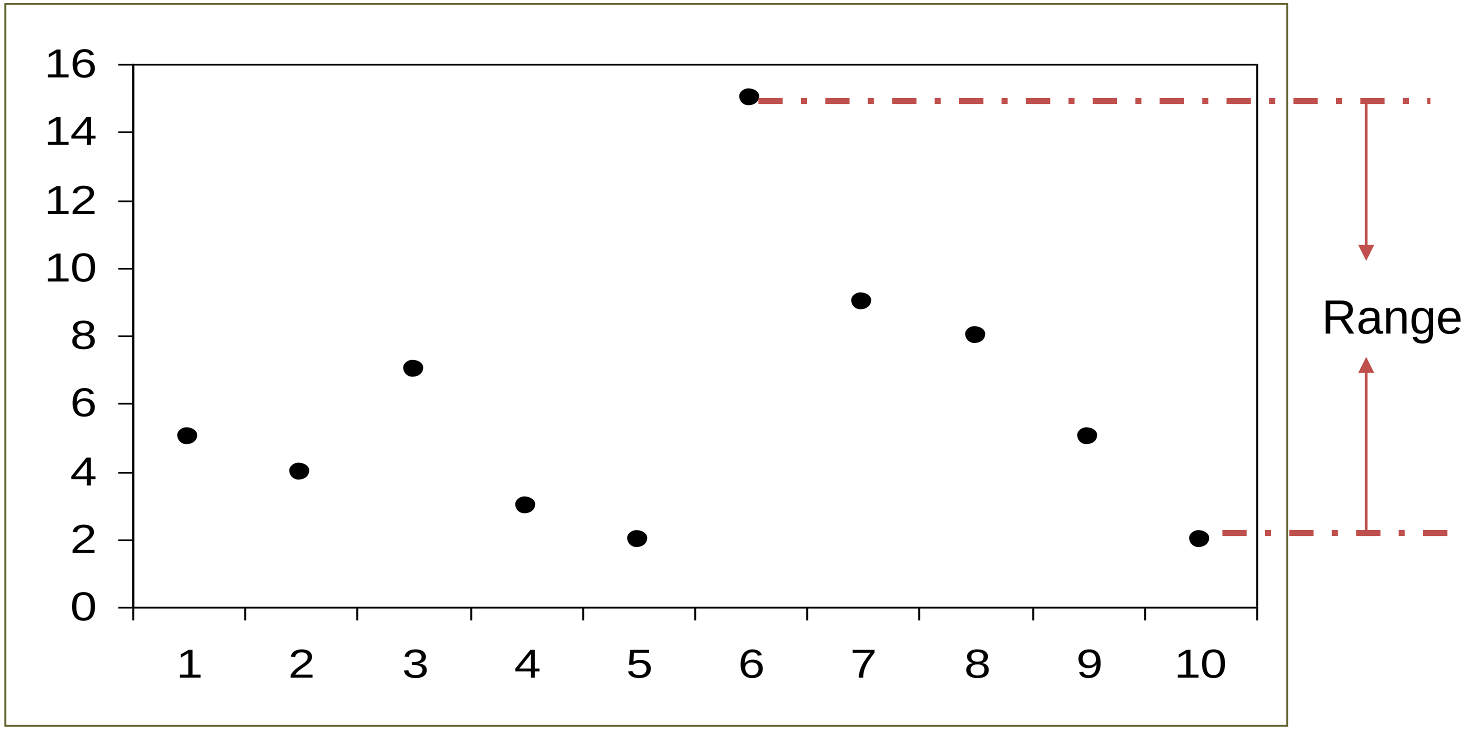
FUNDAMENTALS OF STATISTICS

Summarization of sample data : Measure of Variation or dispersion

Sample Range: Issues

It depends only on extreme values

Hence affected by outliers



Summarization of sample data : Measure of Variation or dispersion

Sample Standard Deviation: Example:

5	4	7	3	2
15	9	8	5	2

Step 1:

Calculate Average

Average = 6

Step 2:

Take deviations from Mean

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

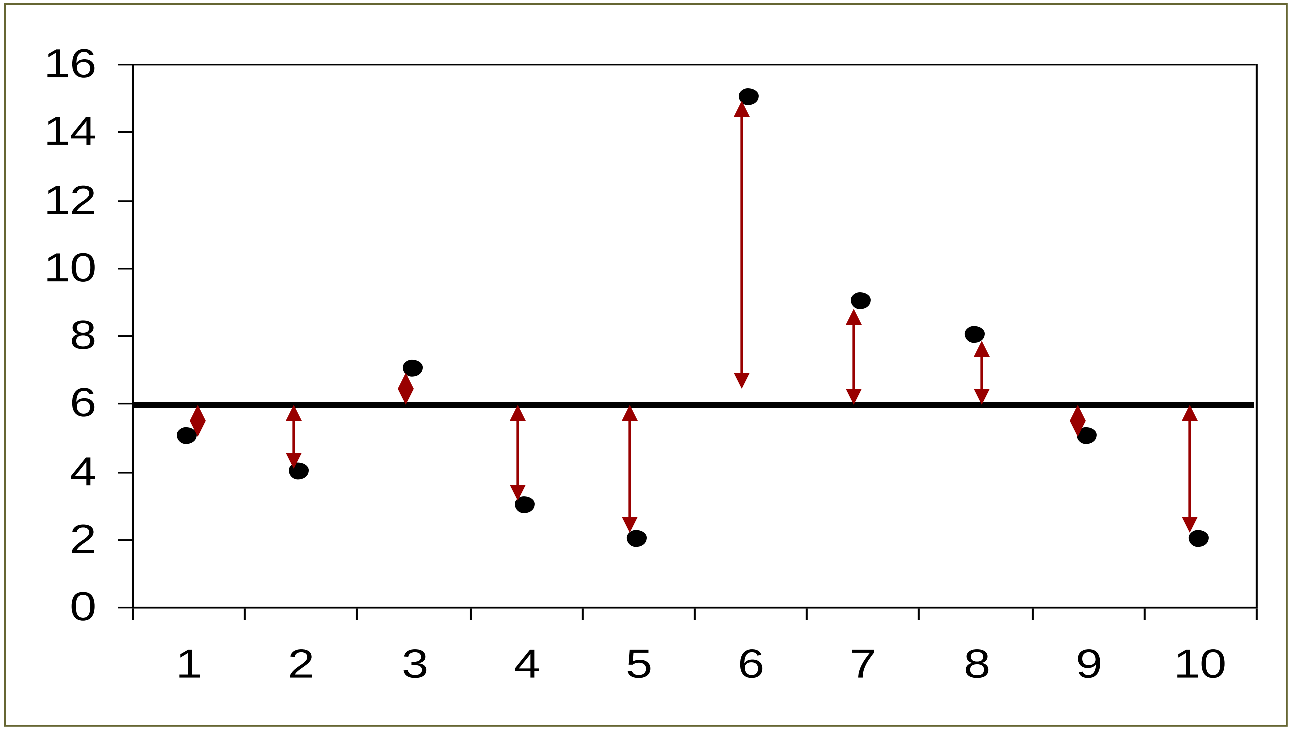
FUNDAMENTALS OF STATISTICS

Summarization of sample data : Measure of Variation or dispersion

Sample Standard Deviation: Example:

Step 2:

Take deviations from Mean



Summarization of sample data : Measure of Variation or dispersion

Sample 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 of squares = 142

Step 4:

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

$= \sqrt{(142 / (10 - 1))}$

$= \sqrt{15.77} = 3.972$

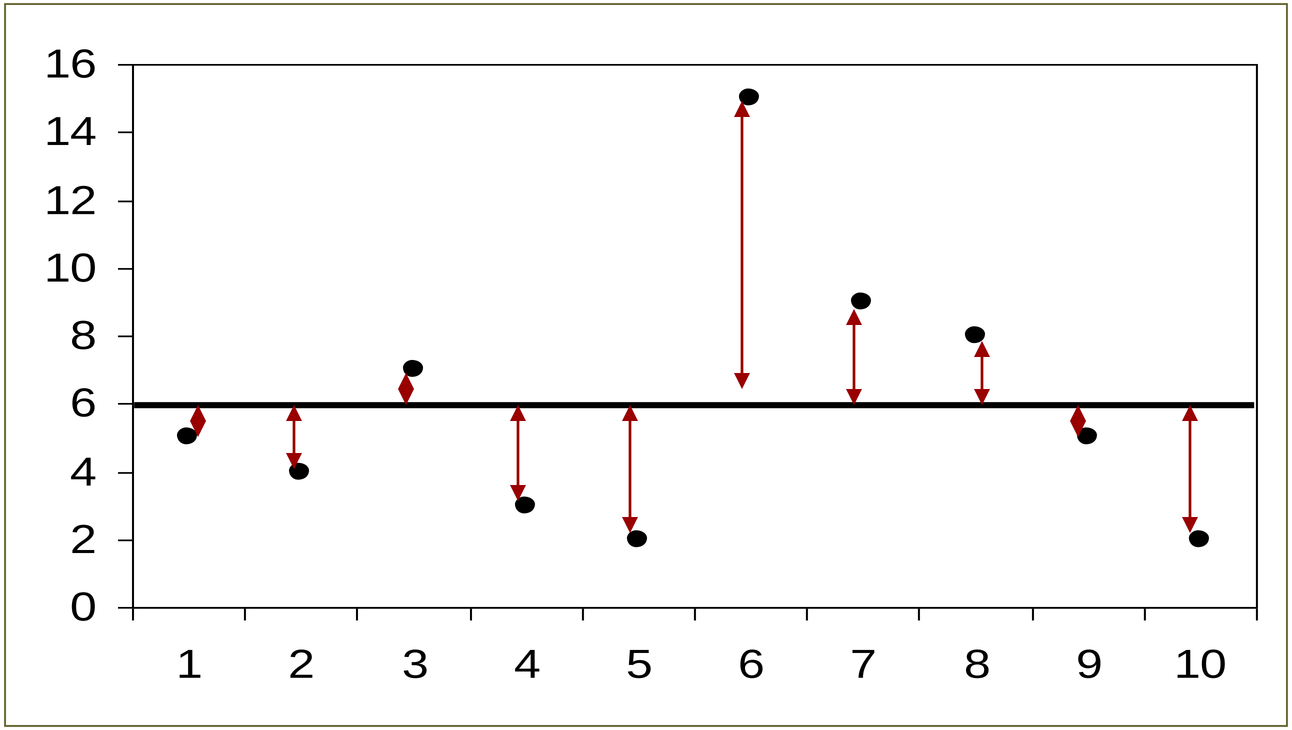
FUNDAMENTALS OF STATISTICS

Summarization of sample data : Measure of Variation or dispersion

Sample Standard Deviation: Interpretation

Square root of the average squared deviation from average

Indicates on an average how much each value is away from the average



Sample Standard Deviation: Credit Card usage data

Month	Credit Card Expenses	Month	Credit Card Expenses
1	55	11	63
2	65	12	55
3	59	13	61
4	59	14	61
5	57	15	57
6	61	16	59
7	53	17	61
8	63	18	57
9	59	19	59
10	57	20	63

Frequency Table

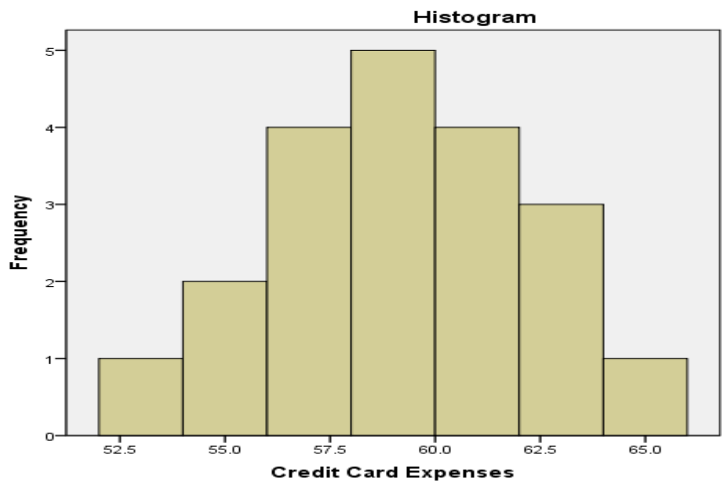
Count of frequency of a variable in a given range/ observation and presented in tabular form.

FUNDAMENTALS OF STATISTICS

Frequency Table: Credit Card usage data

Values	Count	Percent	Cumulative Percent
53	1	5	5
55	2	10	15
57	4	20	35
59	5	25	60
61	4	20	80
63	3	15	95
65	1	5	100
Total	20	100	

Histogram: Graphical representation of frequency table



Exercise: The data of 30 customers on credit card usage in INR1000, gender (1: male, 2: female) and whether they have done shopping or banking (1: yes , 2: no) with credit card are given.

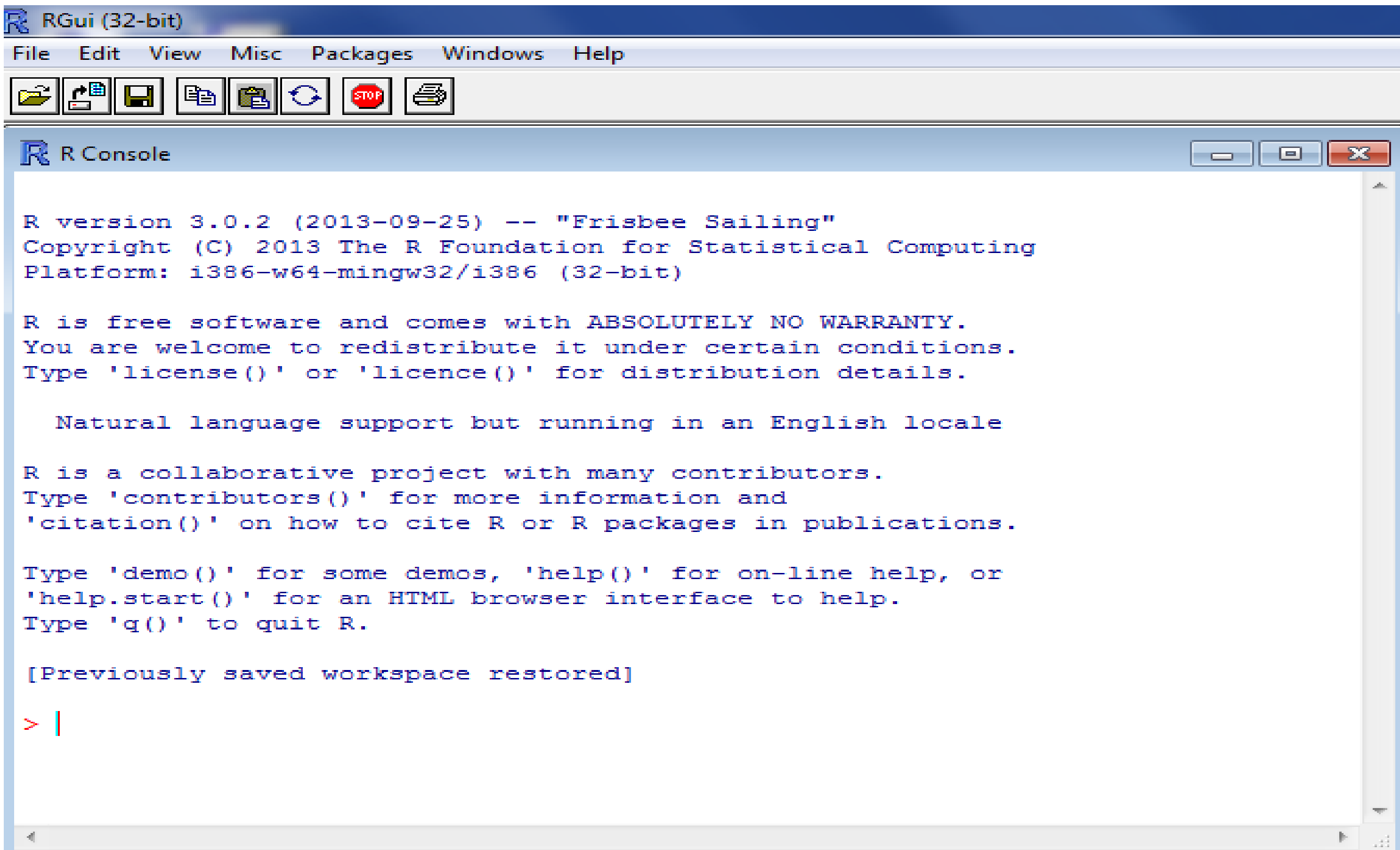
1. Summarize and interpret the credit card usage?
2. How the credit card usage vary with gender?
3. How the credit card usage pattern vary with those who do shopping with credit card and those who don’t do shopping?
4. How the credit card usage pattern vary with those who do banking with credit card and those who don’t do banking?

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

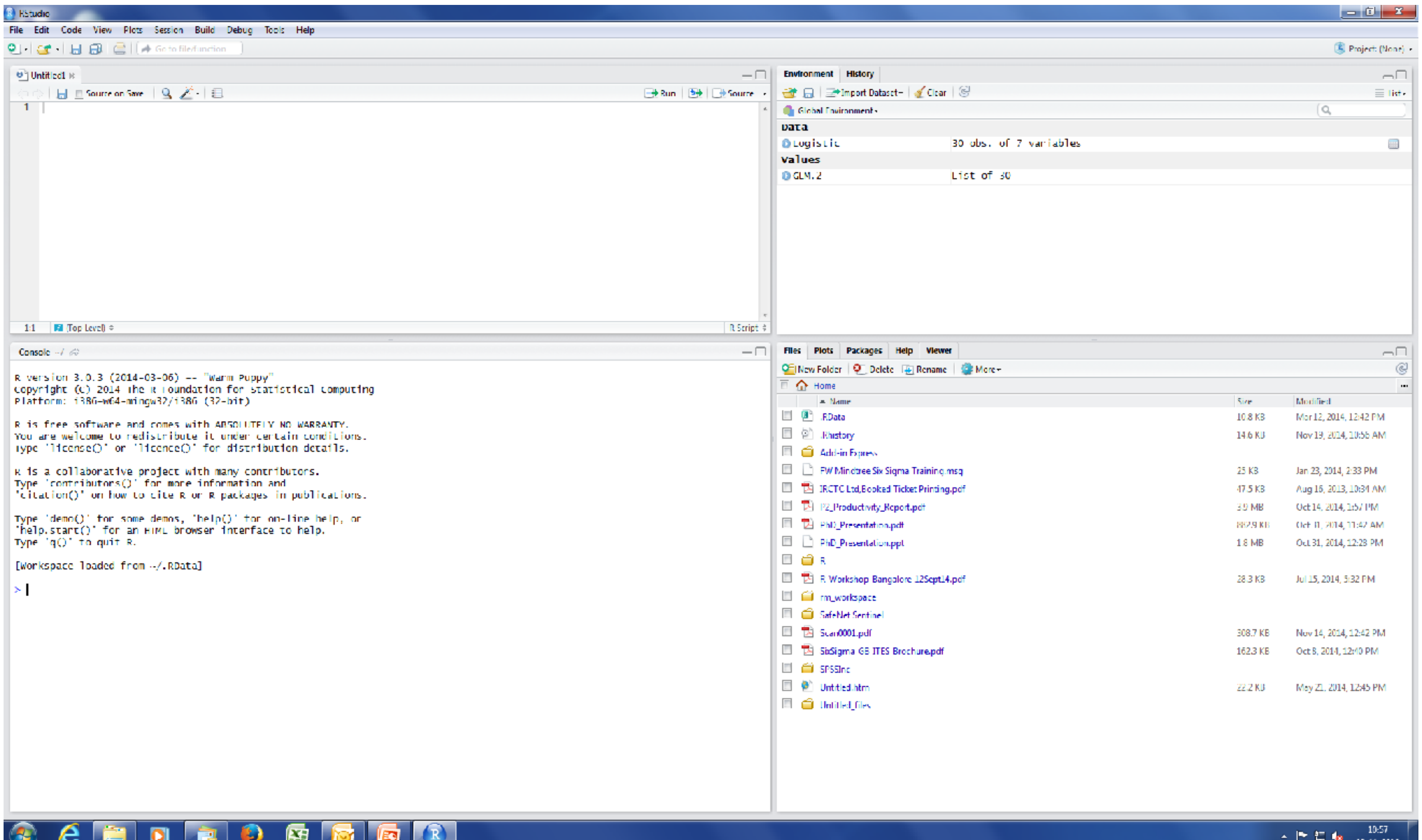
3. Double click on the R icon in the desktop and R window will open



R INSTALLATION

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

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



DESCRIPTIVE STATISTICS
using R

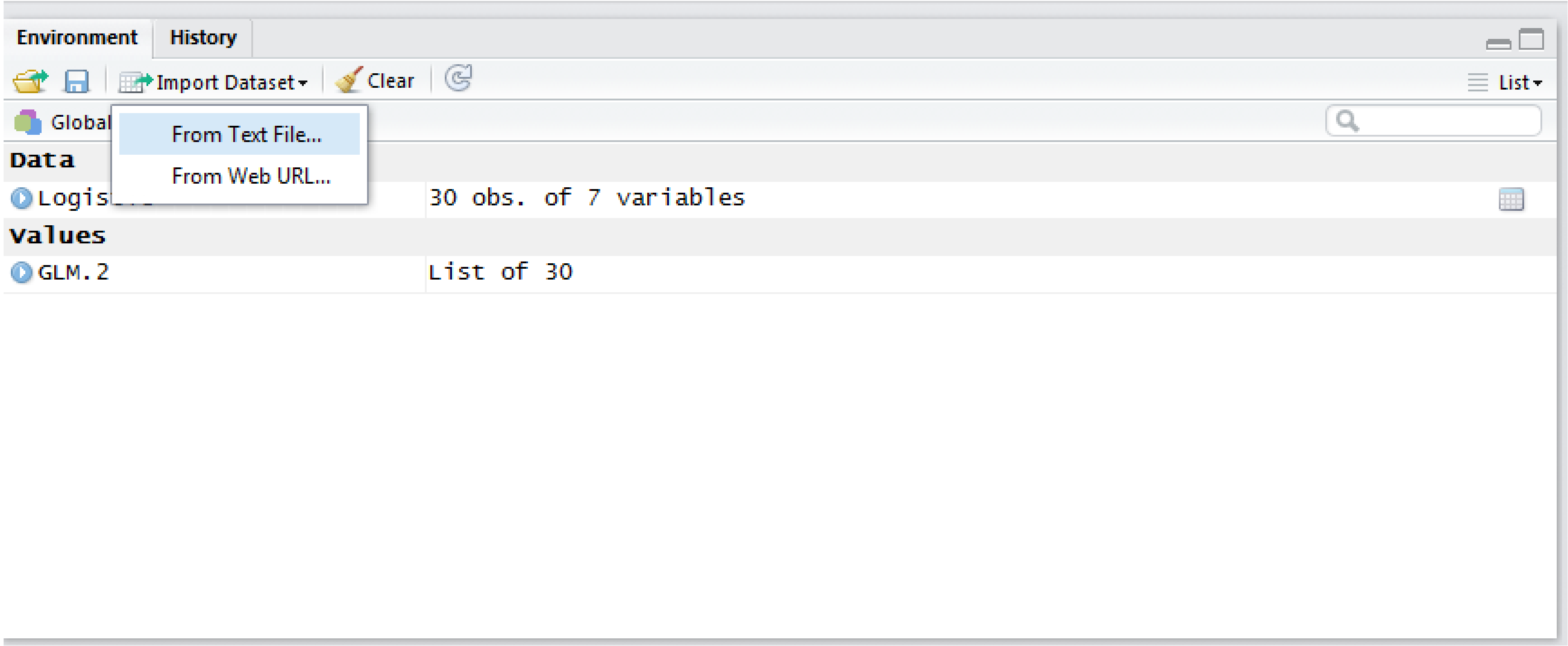
DESCRIPTIVE STATISTICS

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

a. Read the dataset to R studio
b. Compute mean, median minimum, maximum, range, variance, standard deviation, skewness, kurtosis and quantiles of Credit Card Expenses
c. Compute default summary of Credit Card Expenses
d. Draw Histogram of Credit Card Expenses

DESCRIPTIVE STATISTICS

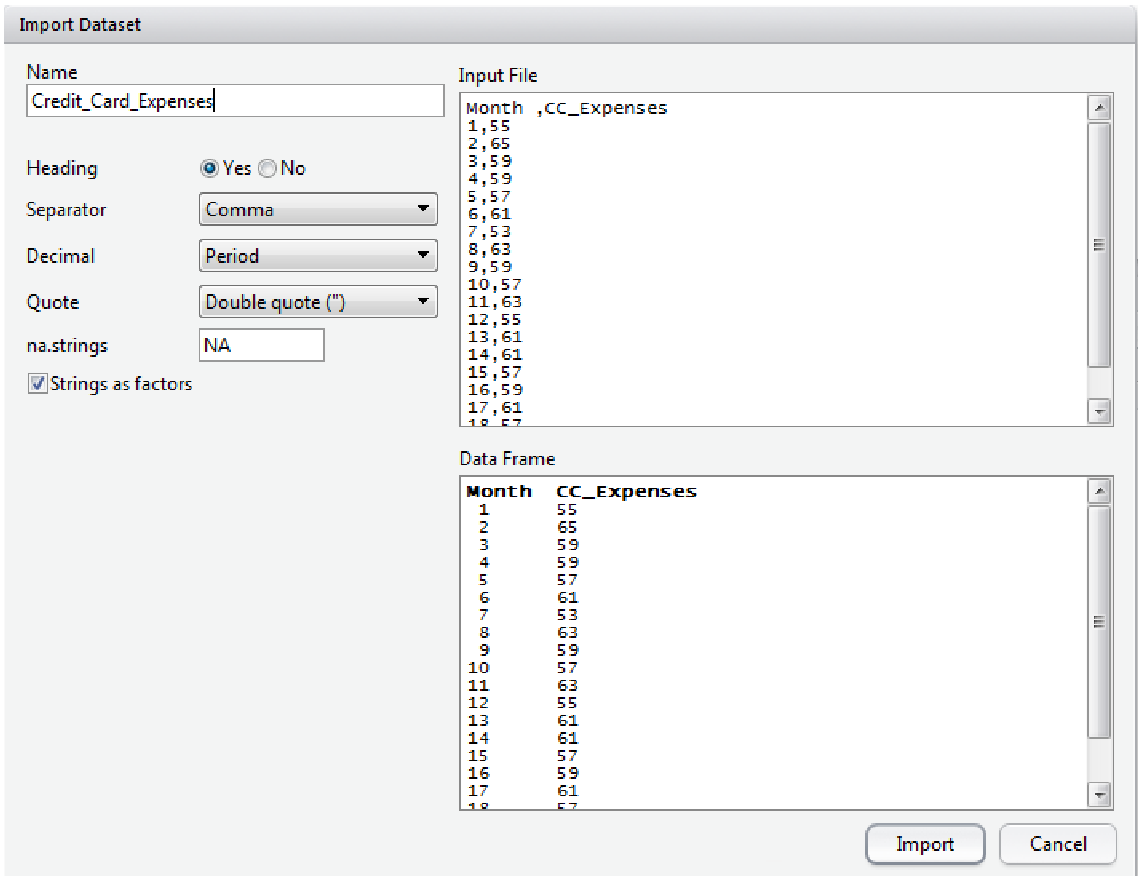
Reading a csv file to R Studio



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

```
> Credit_Card_Expenses <- read.csv("D:/SQC/DataSets/Credit_Card_Expenses.csv")
```

To change the name of the data set to : mydata

```
> mydata = Credit_Card_Expenses
```

To display the contents of the data set

```
> print(mydata)
```

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

Example: Read CC_Expenses to CC

```
>CC = mydata$CC_Expenses
```

DESCRIPTIVE STATISTICS

Reading data from MS Excel formats to R Studio

Format	Code
Excel	library(xlsx) mydata <- read.xlsx("c:/myexcel.xlsx", “Sheet1”)

DESCRIPTIVE STATISTICS

Reading data from databases to R Studio

Function	Description
odbcConnect(*dsn*, uid="", pwd="")	Open a connection to an ODBC database
sqlFetch(*channel*, *sqtable*)	Read a table from an ODBC database into a data frame
sqlQuery(*channel*, *query*)	Submit a query to an ODBC database and return the results
sqlSave(*channel*, *mydf*, tablename = *sqtable*, append = *FALSE*)	Write or update (append=True) a data frame to a table in the ODBC database
sqlDrop(*channel*, *sqtable*)	Remove a table from the ODBC database
close(*channel*)	Close the connection

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

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

DESCRIPTIVE STATISTICS

Descriptive Statistics

Computation of descriptive statistics for variable CC

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

DESCRIPTIVE STATISTICS

Descriptive Statistics

Function	Code
Quantile	> quantile(CC)

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

Function	Code
Summary	>summary(CC)

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

DESCRIPTIVE STATISTICS

Descriptive Statistics

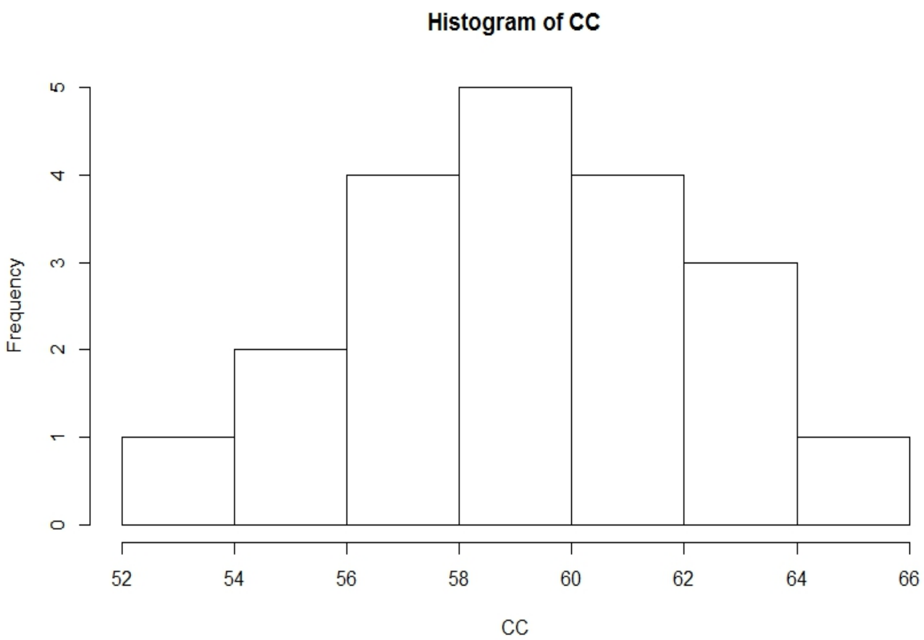
Function	Code
describe	> libray(psych) > describe(CC)

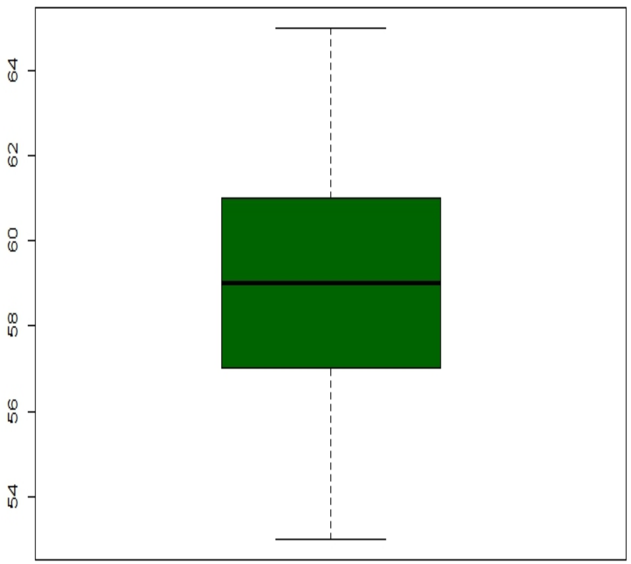
Output	
Statistics	Values
n	20
mean	59.2
sd	3.11
median	59
trimmed	59.25
mad	2.97
min	53
max	65
range	12
skew	-0.08
kurtosis	-0.85
se	0.69

DESCRIPTIVE STATISTICS

Graphs

Graph Type	Code
Histogram	> hist(CC)
Histogram colour ("Blue")	> hist(CC,col="blue")
Dot plot	> dotchart(CC)
Box plot	> boxplot(CC)
Box plot colour	> boxplot(CC, col="dark green")





DESCRIPTIVE STATISTICS

Exercise 2: The data of 30 customers on credit card usage in INR1000, sex (1: male, 2: female) and whether they have done shopping or banking (1: yes , 2: no) with credit card are given in file CC_Expenses_Exercise.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 Credit Card Usage
d. Convert the variables sex, banking & shopping to categorical (factor)
e. Check whether the average usage varies with sex?
f. Check whether the average credit card usage vary with those who do shopping with credit card and those who don't do shopping?
g. Check whether the average credit card usage vary with those who do banking with credit card and those who don't do banking?
h. Compute the aggregate average of usage with sex & shopping?
i. Compute the aggregate average of usage with all three factors?

DESCRIPTIVE STATISTICS

Exercise 2: The data of 30 customers on credit card usage in INR1000, sex (1: male, 2: female) and whether they have done shopping or banking (1: yes , 2: no) with credit card are given in file CC_Expenses_Exercise.csv.

Reading dataset to variable: mydata

```
>mydata = CC_Expenses_Exercise
```

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,"D:/SQC/DataSets/mynewdata.csv")
```

DESCRIPTIVE STATISTICS

Exercise 2: The data of 30 customers on credit card usage in INR1000, sex (1: male, 2: female) and whether they have done shopping or banking (1: yes , 2: no) with credit card are given in file CC_Expenses_Exercise.csv.

Reading variable Credit_Card_Usage to a new variable: CC

```
> CC = mydata$Credit.Card.usage
```

Computing descriptive statistics for variable : CC

```
> summary(CC)
```

Minimum	Q1	Median	Mean	Q3	Maximum
20	30	55	66	90	150

DESCRIPTIVE STATISTICS

Exercise 2: The data of 30 customers on credit card usage in INR1000, sex (1: male, 2: female) and whether they have done shopping or banking (1: yes , 2: no) with credit card are given in file CC_Expenses_Exercise.csv.

Converting variables sex, shopping & banking to factors

```
> sex = factor(mydata$sex)
> banking = factor(mydata$Banking)
> shopping = factor(mydata$Banking)
```

Computing average credit card usage for different sex

```
> CC_sex = aggregate(CC,by=list(sex),FUN = mean)
```

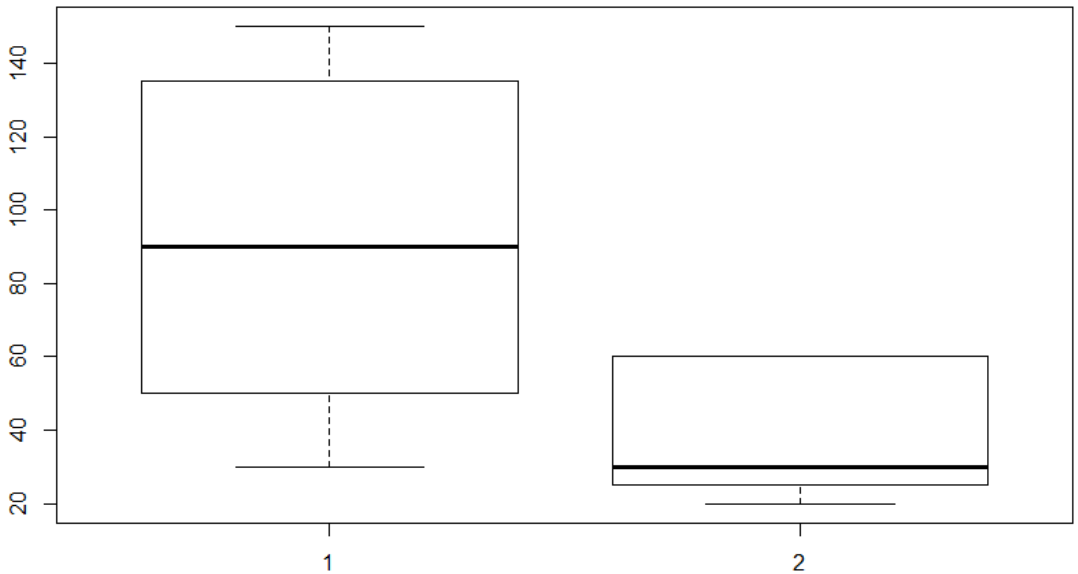
Group	Sex	Average Credit Card Usage
1	Male	93.33333
2	Female	38.66667

DESCRIPTIVE STATISTICS

Exercise 2: The data of 30 customers on credit card usage in INR1000, sex (1: male, 2: female) and whether they have done shopping or banking (1: yes , 2: no) with credit card are given in file CC_Expenses_Exercise.csv.

Box plot of Credit Card usage by sex

```
> boxplot(CC~sex)
```



DESCRIPTIVE STATISTICS

Exercise 2: The data of 30 customers on credit card usage in INR1000, sex (1: male, 2: female) and whether they have done shopping or banking (1: yes , 2: no) with credit card are given in file CC_Expenses_Exercise.csv.

Computing aggregate average of credit card usage for different sex and shopping

```
> CC_sex_bank = aggregate(CC, by = list(sex, banking), FUN = mean)
```

Sex	Banking	Average Credit Card Usage
Male	Yes	115.00000
Female	Yes	40.00000
Male	No	68.57143
Female	No	38.57143

DESCRIPTIVE STATISTICS

Exercise 2: The data of 30 customers on credit card usage in INR1000, sex (1: male, 2: female) and whether they have done shopping or banking (1: yes , 2: no) with credit card are given in file CC_Expenses_Exercise.csv.

Computing aggregate average of credit card usage by 3 factors

```
CC_Aggregate = aggregate(CC, by = list(sex, banking, shopping), FUN = mean)
```

Sex	Banking	Shopping	Average Credit Card Usage
Male	Yes	Yes	130.00000
Female	Yes	Yes	40.00000
Male	No	Yes	62.00000
Female	No	Yes	48.00000
Male	Yes	No	70.00000
Male	No	No	85.00000
Female	No	No	33.33333

DESCRIPTIVE STATISTICS

Exercise 2: The data of 30 customers on credit card usage in INR1000, sex (1: male, 2: female) and whether they have done shopping or banking (1: yes , 2: no) with credit card are given in file CC_Expenses_Exercise.csv.

Computing aggregate summary of credit card usage by 3 factors

```
> CC_Aggregate = aggregate(CC, by = list(sex, banking, shopping), FUN = summary)
```

Sex	Banking	Shopping	Credit Card Expenses					
			Minimum	Q1	Median	Mean	Q3	Maximum
Male	Yes	Yes	90	130	135	130	140	150
Female	Yes	Yes	40	40	40	40	40	40
Male	No	Yes	30	40	40	62	50	150
Female	No	Yes	30	30	60	48	60	60
Male	Yes	No	50	60	70	70	80	90
Male	No	No	80	82.5	85	85	87.5	90
Female	No	No	20	20	30	33.33	40	60

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)
```

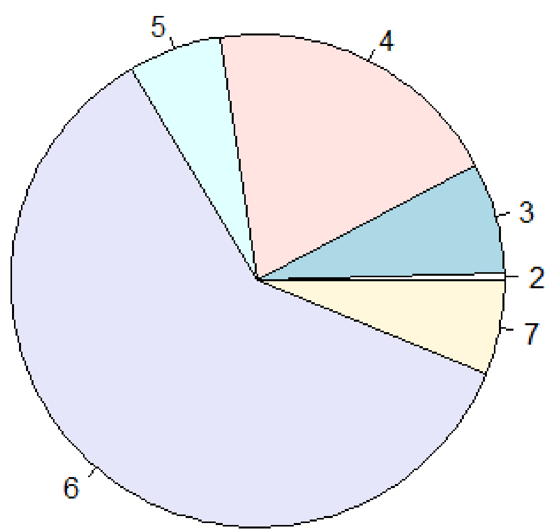
Rating	Frequency
2	1
3	13
4	35
5	11
6	108
7	11

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)
```

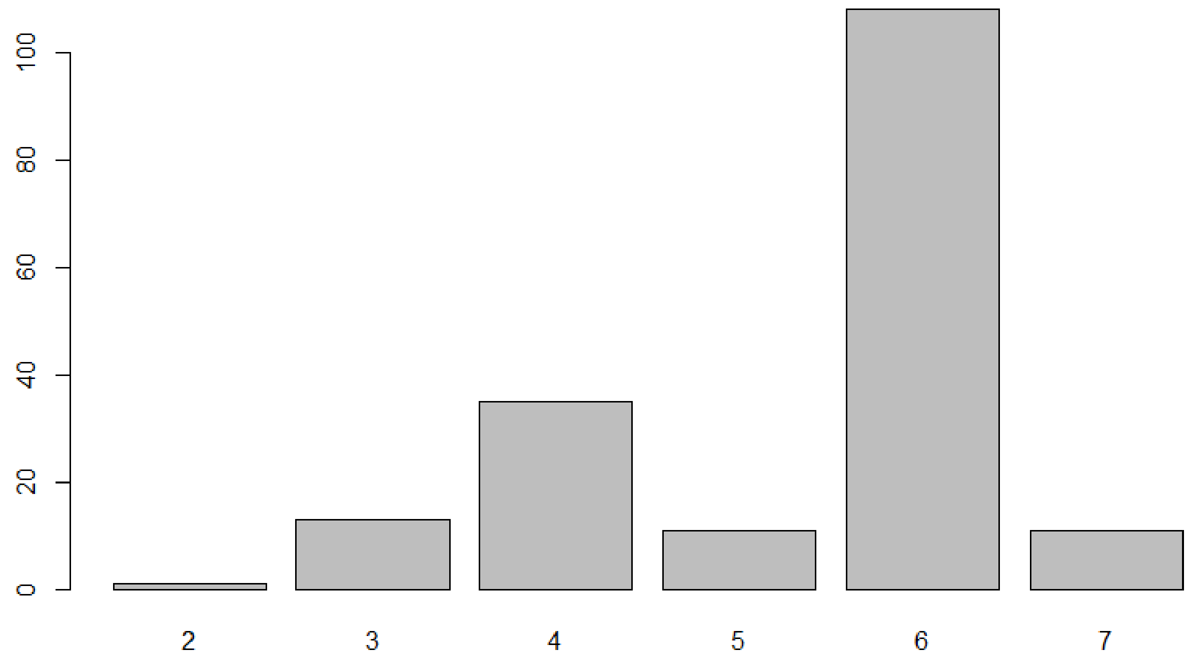


DESCRIPTIVE STATISTICS

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)
```



DATA PREPROCESSING

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

Missing Value Handling

Example: Suppose a telecom company wants to analyze the performance of its circles based on the following parameters

1. Current Month's Usage
2. Last 3 Month's Usage
3. Average Recharge
4. Projected Growth

The data set is given in next slide. Read this data set to RapidMiner

Missing Value Handling

Example:

Circle wise Data

Read data and variables to R

```
> mydata = Missing_Values_Telecom
> cmusage = mydata[,2]
> l3musage = mydata[,3]
> avrecharge = mydata[,4]
```

SL No.	Current Month's Usage	Last 3 Month's Usage	Average Recharge	Projected Growth	Circle
1	5.1	3.5	99.4	99.2	A
2	4.9	3	98.6	99.2	A
3		3.2		99.2	A
4	4.6	3.1	98.5	9..2	A
5	5		98.4	99.2	A
6	5.4	3.9	98.3	99.4	A
7	7	3.2	95.3	98.4.	B
8	6.4	3.2	95.5	98.5	B
9	6.9	3.1	95.1	98.5	B
10		2.3	96	98.3	B
11	6.5	2.8	95.4	98.5	B
12	5.7		95.5	98.3	B
13	6.3	3.3		98.6	B
14	6.7	3.3	94.3	97.5	C
15	6.7	3	94.8	97.3	C
16	6.3	2.5	95	98.9	C
17		3	94.8	98	C
18	6.2	3.4	94.6	97.3	C
19	5.9	3	94.9	98.8	C

Missing Value Handling

Option 1: Discard all records with missing values

```
>newdata = na.omit(mydata)

>write.csv(newdata,"E:/ISI_Mumbai/newdata.csv")
```

SL.No.	Current.Month.s.Usage	Last.3.Month.s.Usage	Average.Recharge	Projected.Growth	Circle
1	5.1	3.5	99.4	99.2	A
2	4.9	3	98.6	99.2	A
4	4.6	3.1	98.5	9..2	A
6	5.4	3.9	98.3	99.4	A
7	7	3.2	95.3	98.4.	B
8	6.4	3.2	95.5	98.5	B
9	6.9	3.1	95.1	98.5	B
11	6.5	2.8	95.4	98.5	B
14	6.7	3.3	94.3	97.5	C
15	6.7	3	94.8	97.3	C
16	6.3	2.5	95	98.9	C
18	6.2	3.4	94.6	97.3	C
19	5.9	3	94.9	98.8	C

Missing Value Handling

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

Replacing the missing values with men

Compute the means excluding ghe missing values

```
>cmusage_mean = mean(cmusage, na.rm = TRUE)
>l3musage_mean = mean(l3musage_mean, na.rm = TRUE)
> l3musage_mean = mean(l3musage, na.rm = TRUE)
> avrecharge_mean = mean(avrecharge, na.rm = TRUE)
```

Replace the missing values with mean

```
> cmusage[is.na(cmusage)]=cmusage_mean
> l3musage[is.na(l3musage)]= l3musage_mean >
>avrecharge[is.na(avrecharge)]=avrecharge_mean
```

Missing Value Handling

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

Replacing the missing values with men

Replace the missing values with mean

```
> cmusage[is.na(cmusage)]=cmusage_mean
> l3musage[is.na(l3musage)]= l3musage_mean
>avrecharge[is.na(avrecharge)]=avrecharge_mean
```

Making the new file

```
> mynewdata = cbind(cmusage, l3musage, avrecharge, mydata[,5],mydata[,6])

> write.csv(mynewdata, "E:/ISI_Mumbai/mynewdata.csv")
```

Missing Value Handling

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

Replacing the missing values with men

SL No	cmusage	l3musage	avrecharge	Proj Growth	Circle
1	5.1	3.5	99.4	11	1
2	4.9	3	98.6	11	1
3	5.975	3.2	96.14117647	11	1
4	4.6	3.1	98.5	1	1
5	5	3.105882353	98.4	11	1
6	5.4	3.9	98.3	12	1
7	7	3.2	95.3	6	2
8	6.4	3.2	95.5	7	2
9	6.9	3.1	95.1	7	2
10	5.975	2.3	96	5	2
11	6.5	2.8	95.4	7	2
12	5.7	3.105882353	95.5	5	2
13	6.3	3.3	96.14117647	8	2
14	6.7	3.3	94.3	3	3
15	6.7	3	94.8	2	3
16	6.3	2.5	95	10	3
17	5.975	3	94.8	4	3
18	6.2	3.4	94.6	2	3
19	5.9	3	94.9	9	3

DATA MERGING

Exercise: The data of 30 customers on credit card usage in INR1000 is given in CC_Usage.txt. Similarly the user profile namely gender (1: male, 2: female) and whether they have done shopping or banking (1: yes , 2: no) with credit card are given in cc_Profile.csv. Can you merge the two files into a single data set?

Read the files

```
>myprofile = CC_Profile
> myusage = CC_Usage
```

Merge the files by “ID” field

```
>mydata = merge(myprofile, myusage, by = “ID”)
```

DATA APPEND

Exercise: The data on user profile of customers whom are included in the previous mailing campaign is compiled into two files namely classification1.csv and classification2.txt. Can you append the second data set with the first one and store the new data set in a new file?

Read the files

```
>class1 = Classification1
> class2 = Classification2
```

Append class1 with class2

```
>mydata = rbind(class1, class2)
```

TRANSFORMATION / NORMALIZATION

z transform:

Transformed data = (Data – Mean) / SD

Exercise : Normalize the variables in the factor_Analysis_Example.csv ?

Read the files

```
>mydata  = Factor_Analysis_Example
> mydata = mydata[,2:7]
```

Normalize or standardize the variable

```
>mystddata = scale(mydata)
```

RANDOM SAMPLING

Example: Take a sample of size 60 (10%) randomly from the data given in the file bank-data.csv and save it as a new csv file?

Read the files

```
>mydata  = bank-data

> mysample = mydata[sample(1:nrow(mydata), 60, replace = FALSE),]

>write.csv(mysample,"E:/ISI_Mumbai/mysample.csv")
```

RANDOM SAMPLING

Example: Split randomly the data given in the file bank-data.csv into sets namely training (75%) and test (25%) ?

Read the files

```
>mydata  = bank-data

>sample = sample(2, nrow(mydata), replace = TRUE, prob = c(0.75, 0.25))
> sample1 = mydata[sample ==1, ]
> sample2 = mydata[sample ==2,]
```

Fundamentals
of
Probability

FUNDEMENTALS OF PROBABILITY

An Event

An event is one or more of the possible outcomes of doing some things. If we toss a coin, getting a tail is an event, and getting a head is another event.

An Experiment

An experiment is an activity that produces an event.

Tossing a coin, Drawing a card from a deck of cards.

Sample Space

The set of all possible outcomes of an experiment is called the sample space for the experiment.

In a coin toss experiment, sample space is {head and tail}.

FUNDEMENTALS OF PROBABILITY

- Probability is a chance of an event occurring .

- Probability of an event is the ratio of chance favoring the event by total possible event

$$\text{Probability of an event} = \frac{\text{Chances favoring the event}}{\text{Total possible events}}$$

when total possible events are very large.

FUNDEMENTALS OF PROBABILITY

Example

Tossing of a coin is an experiment.

Here,

Sample Space S={head,tail};

Event 1- getting the head;

Event 2- getting the tail;

In tossing of a coin experiment, what is the probability of getting a head????

probability p(getting head)= 1/2

Axioms of Probability

- A function P that assigns areal number P(A) to each event A is a ***probability distribution*** or a ***probability measure*** if it satisfies the following three axioms
  a. $P(A) \geq 0$
  b. $P(\Omega) = 1$
  c. If A_1, A_2, ….∞ are disjoint events, i.e. $A_i \cap A_j = \phi$ where ϕ is the empty set, then $P(\cup A_j) = \Sigma\, P(A_j)$

The axioms of probability provides the theoretical basis and the elementary properties mentioned in the previous slide follows from the axioms

FUNDEMENTALS OF PROBABILITY

Important terms:

Two events are said to be mutually exclusive if one and only one of them can take place at a time.

- In our example of Tossing a Coin only Head or Tail can occur

When a list of the possible events that can result from an experiment includes every possible outcome, the list is said to be collectively exhaustive.

- In our example the list “head and tail” is collectively exhaustive.

When outcome of one event does not influence the outcome of another event, the two events are called independent events.

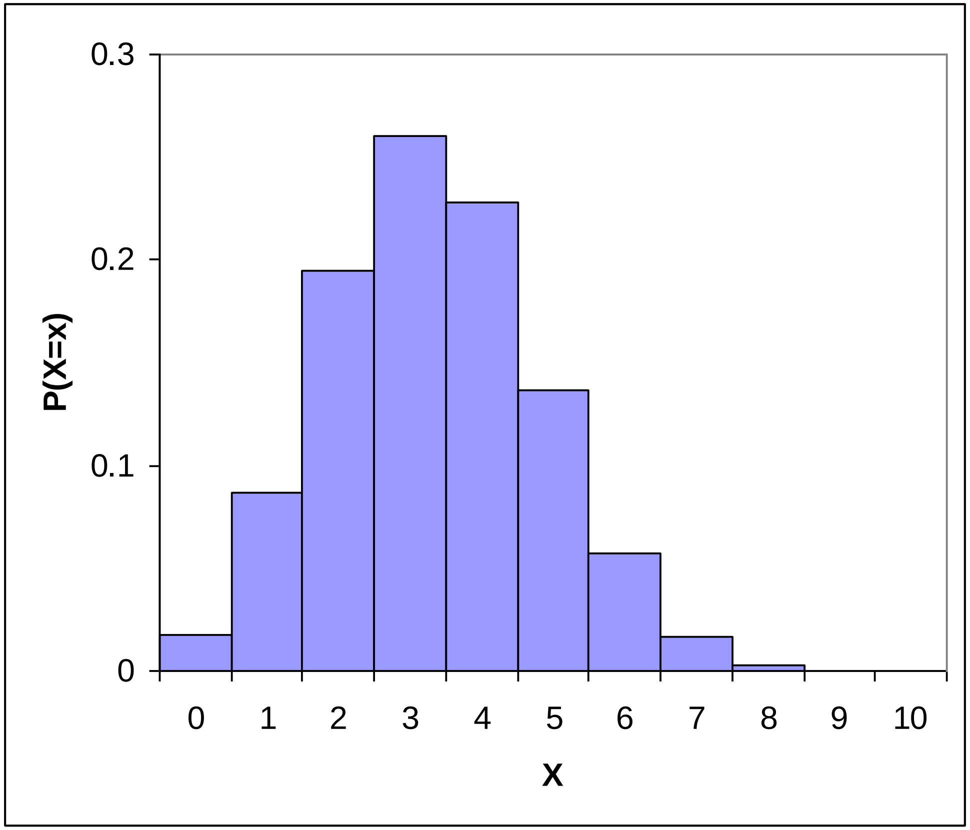
- In our example the outcome of 1st Tossing and 2nd Tossing are independent.

Binomial Distribution

The number of successes in n Bernoulli trials.

Or the sum of n Bernoulli random variables.

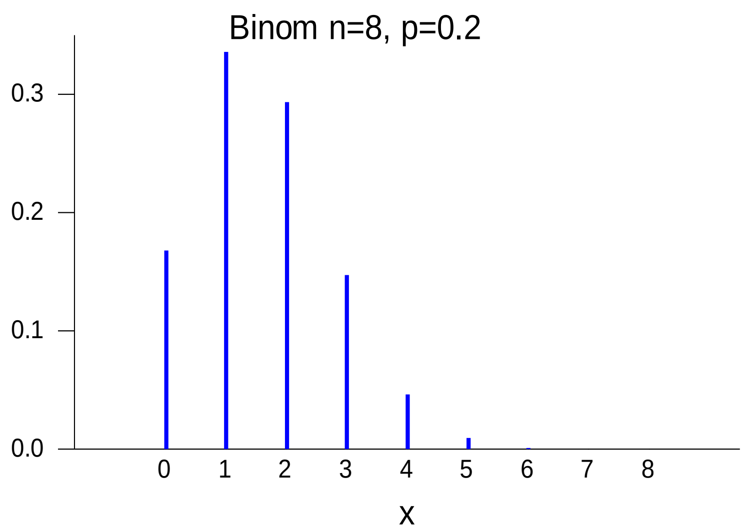
$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

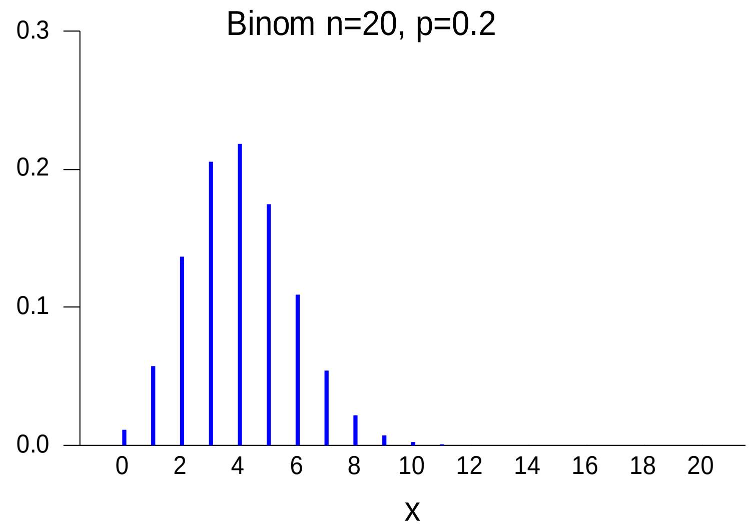


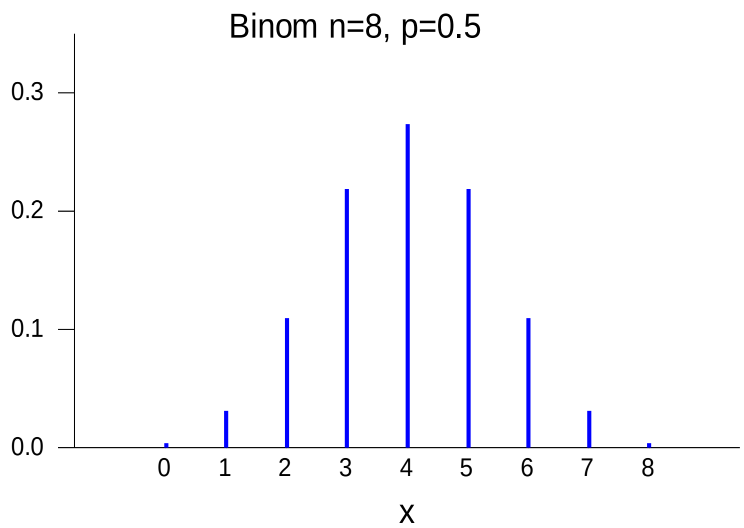
$$E[X] = np$$

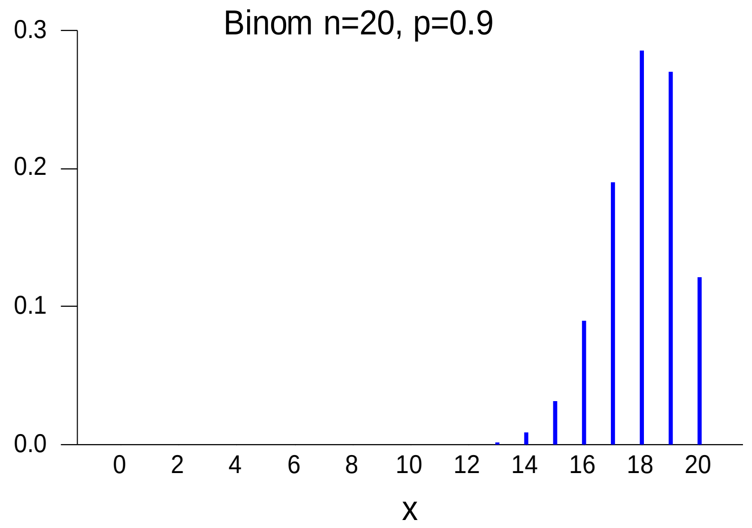
$$Var(X) = np(1-p)$$

Binomial Distribution Plots



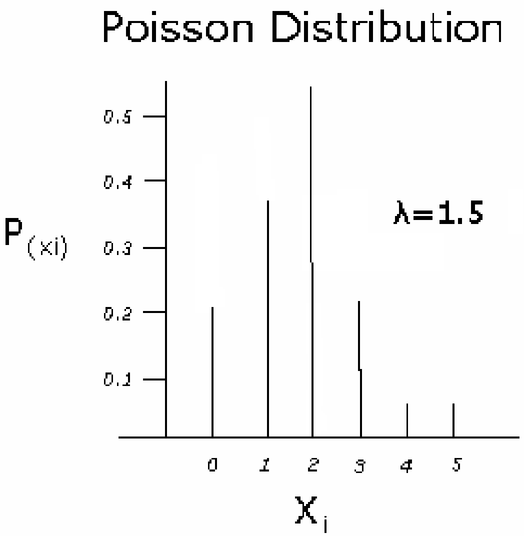






Poisson Distribution

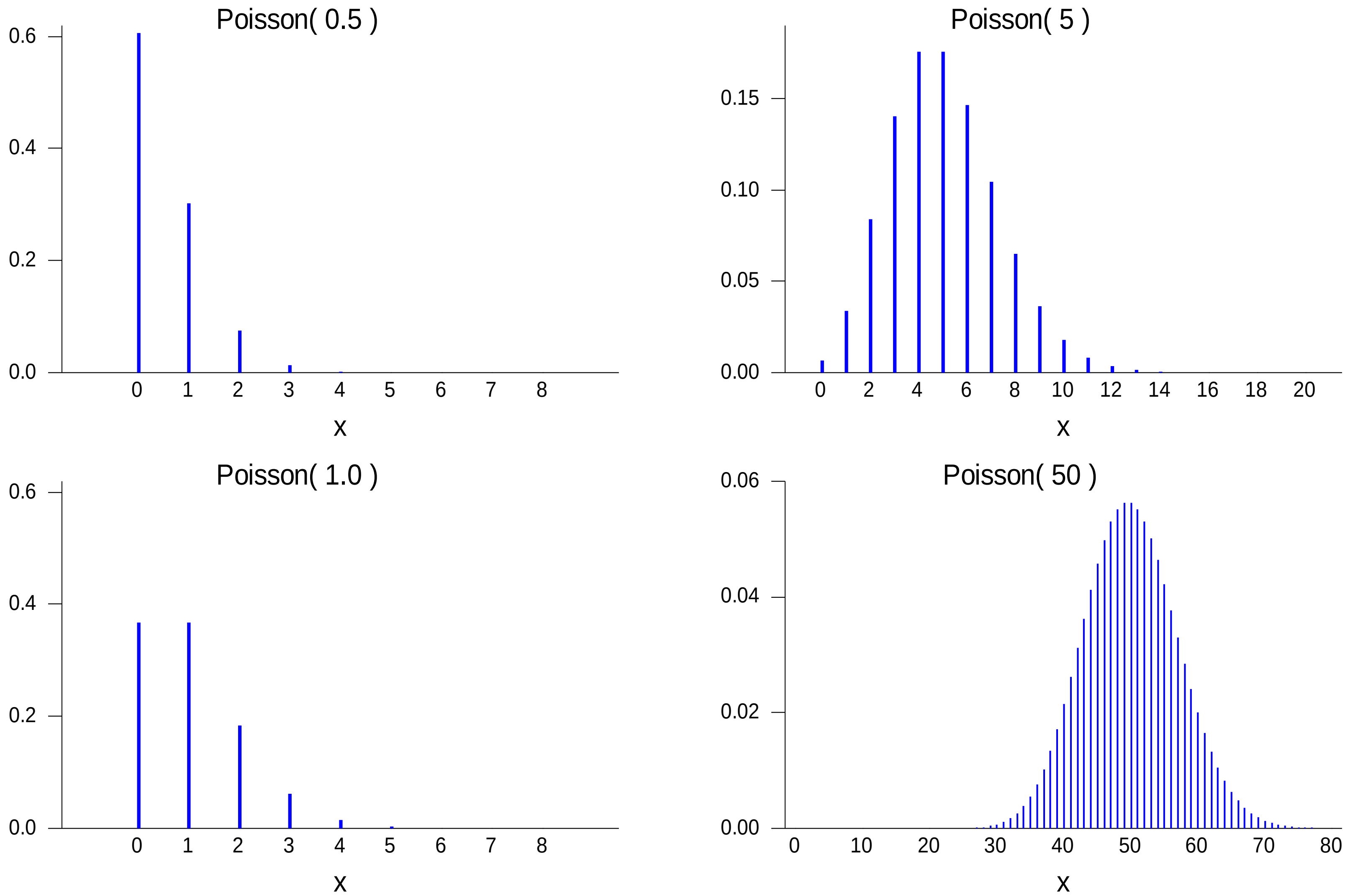
- Poisson distribution also describes discrete data – situations where the random variable can take integer values. Examples are:
  - Number of patients arriving at a physician's office, Number of cars arriving at a toll booth.
- Measures of central tendency and dispersion, for the Poisson distribution
  - Mean = Number of occurrences per interval of time
  - Standard deviation = $\sqrt{mean}$



$$P(x) = \frac{\lambda^x e^{-\lambda}}{x\,!}$$

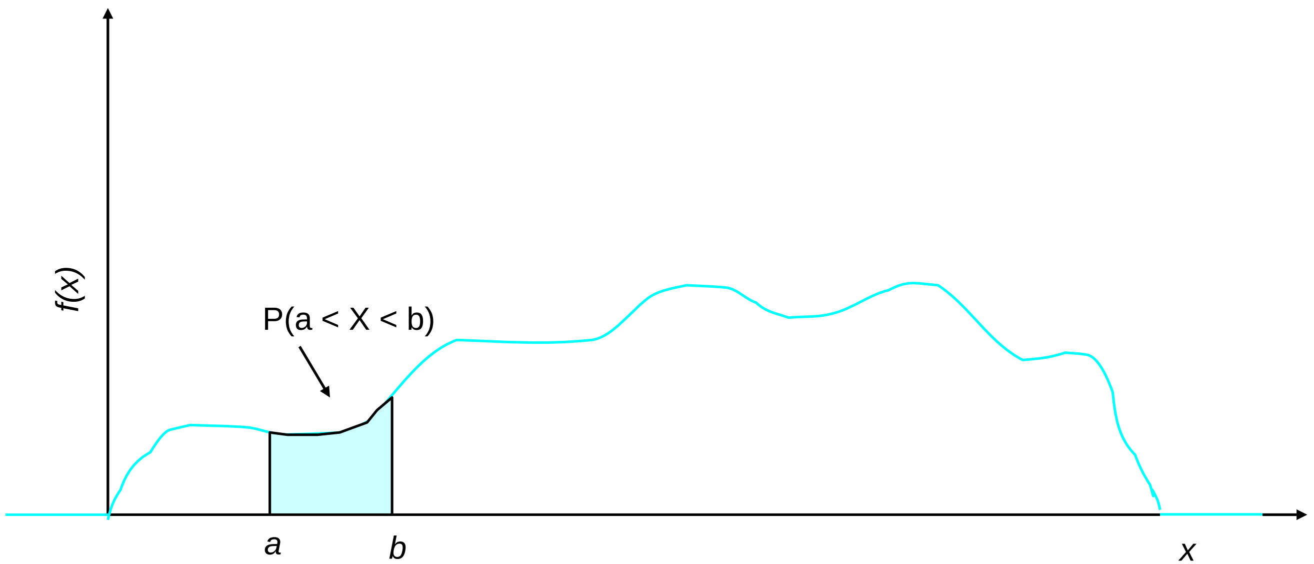
When n>20, or when the number of observations are very large, it has been statistically proven that the Poisson distribution becomes a very good approximation of the binomial distribution.

Poisson Distribution Plots



FUNDEMENTALS OF PROBABILITY

Probability Density Function



Density function of loading on a long, thin beam

For a continuous random variable X, a probability density function is a function such that

(1) $f(x) \geq 0$

(2) $\int_{-\infty}^{\infty} f(x)dx = 1$

(3) $P(a \leq X \leq b) = \int_a^b f(x)dx = area\ under\ f(x)\ from\ a\ to\ b\ for\ any\ a\ and\ b$

Uniform Distribution

A continuous random variable X with probability density function

$$f(x) = 1/(b-a), \qquad a \le x \le b$$

has a **continuous uniform distribution**

The mean and variance of a continuous uniform random variable X over a $\le$ x $\le$ b are

$$\mu = E(X) = (a+b)/2 \qquad and \qquad \sigma^2 = V(X) = (b-a)^2/12$$

Applications:

- Generating random sample
- Generating random variable

Normal Distribution

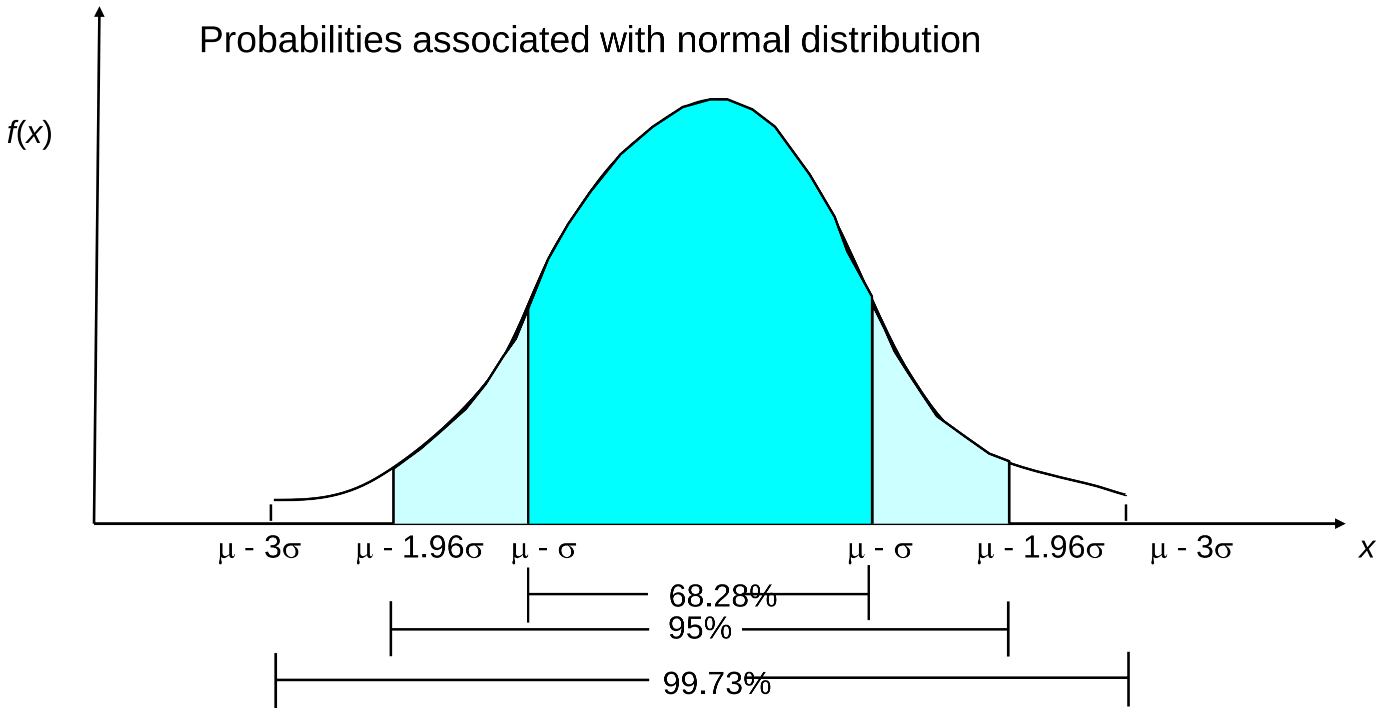
A random variable *X* with probability density function

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \qquad for \; -\infty < x < \infty$$

has a normal distribution with parameters μ, where -∞ < μ < ∞ , and σ > 0. Also,

$$E(X) = \mu \quad and \quad V(X) = \sigma^2$$

Probabilities associated with normal distribution



Standard Normal

A normal random variable with $\mu = 0$ and $\sigma^2 = 1$ is called a standard normal random variable. A standard normal random variable is denoted as *Z*.

$$\Phi(z) = P(Z \leq z)$$

The CDF of a standard normal random variable is denoted as

Standardization

If *X* is a normal random variable with $E(X) = \mu$ and $V(X) = \sigma^2$, then the random variable

$$Z = \frac{X - \mu}{\sigma}$$

is a normal random variable with $E(Z) = 0$ and $V(Z) = 1$. That is , *Z* is a standard normal random variable.

Standardization

Suppose X is a normal random variable with mean μ and variance σ^2 . Then,

$$P(X \leq x) = P\left(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right) = P(Z \leq z)$$

where,

Z is a **standard normal random variable**, and
$z = (x - \mu)/\sigma$ is the z-value obtained by **standardizing** X.

Applications:

- Modeling errors
- Modeling grades
- Modeling averages

Exponential Distribution

The random variable X that equals the distance between successive counts of a Poisson process with mean $\lambda > 0$ has an exponential distribution with parameter λ. The probability density function of X is

$$f(x) = \lambda e^{-\lambda x}, \qquad for\ 0 \le x < \infty$$

If the random variable X has an exponential distribution with parameter λ , then

$E(X) = 1/\lambda$ and $V(X) = 1/\lambda^2$

Lack of Memory Property

For an exponential random variable *X*,

$$P(X < t_1 + t_2 \mid X > t_1) = P(X < t_2)$$

Applications:

- Models random time between failures
- Models inter-arrival times between customers

R Functions

Distribution	Function	Description
Normal	dnorm(*x*)	normal density function (by default m=0 sd=1)
	pnorm(*q*)	cumulative normal probability for q
	qnorm(*p*)	Inverse Normal (quantile)
	rnorm(*n*, m=0,sd=1)	n random normal deviates with mean m
Binomial	dbinom(*x*, *size*, *prob*)	binomial density function
	pbinom(*q*, *size*, *prob*)	binomial cumulative density function
	qbinom(*p*, *size*, *prob*)	inverse binomial (quantile)
	rbinom(*n*, *size*, *prob*)	random numbers from binomial distribution
Poisson	dpois(*x*, *lamda*)	poisson density function
	ppois(*x*,*lamda*)	poisson cumulative density function
	qpois(*p*, *lamda*)	inverse poisson(quantile)
	rpois(*n*, *lamda*)	random numbers from binomial distribution

DISTRIBUTIONS

Prefix d for density function, p for cumulative, q for inverse and r for random number generation

R Function	Distribution	Parameters		
beta	beta	shape1,	shape2	
binom	inomial	Sample size	probability	probability
cauchy	Cauchy	location,	scale	
exp	exponential	rate (lamda)		
chisq	chi-squared	x	df	
f	Fisher's	F	df1,	Df2
gamma	gamma	shape		
geom	Geometric	probability		
hyper	hypergeometric	m,	n,	k
lnorm	lognormal	mean,	sd	
logis	Logistic	location,	scale	
nbinom	negative	binomial	size,	Probability
norm	normal	mean,	sd	
pois	Poisson	mean		
t	t	probability	df	
unif	uniform	minimum,	maximum	
weibull	Weibull	shape		

Binomial Distribution

Exercise 1: An electronic product contains 40integrated circuits. The probability that any integrated circuit is defective is 0.01 and the integrated circuits are independent. The product operated only if there are no defective integrated circuits. What is the probability that the product operates?

R code

```
> n = 40
> p = 0.01
> dbinom(0,n,p) or
 > pbinom(0,n,p)
```

Probability that the product operates = 0.6689718

Binomial Distribution

Exercise 2: Because not all passengers show up for their reserved seat, an airline sells 125 tickets for a flight that holds only 120 passengers. The probability that a passenger will show up is 0.9.

a. What is the probability that every passenger who show up will not get a seat?
b. What is the probability that the flight departs with empty seats?

Poisson Distribution

Exercise 1: The number of tickets arrives in a application support centre is Poisson distributed. Suppose the average number of tickets arrives per hour is 10.

a. What is the probability that exactly 5 tickets arrives in one hour?
b. What is the probability that 3 or less tickets arrives in one hour?
c. What is the probability that 15 or more tickets arrives in two hour?
d. What is the probability that 5 or more tickets arrives in half an hour?

R code

```
> mean 5
> dpois(5,10)
```

Probability that exactly 5 tickets arrives in one hour = 0.03783327

Normal Distribution

Exercise 1: The compressive strength of samples of cement can be modelled by a normal distribution with mean of 6000 kg/cm^2 and a standard deviation of 100 kg/cm^2 .

a. What is the probability that a sample's strength is less than 6250 kg/cm^2?
b. What is the probability that a sample's strength is between 5800 and 5900 kg/cm^2?
c. What strength is exceeded by 95%of the samples?

R code

```
> mean = 6000
> sd = 100
> pnorm(6250,mean,sd)
```

Probability that that a sample's strength is less than 6250 kg/cm^2 = 0.99379

Normal Distribution

Exercise 2: The tensile strength of a paper is modelled by a normal distribution with mean of 35 pounds/inch^2 and a standard deviation of 2 pounds/inch^2.

a. What is the probability that a sample's strength is less than 40 pounds/inch^2?
b. If the specification of tensile strength is not to exceed 35pounds/inch^2, what proportion of the samples is scrapped?

Exercise 3: The reaction time of a driver to visual a stimulus is normally distributed with a mean of 0.4 seconds and standard deviation of 0.05 seconds. Simulate 100 instances of reaction time?

Exponential Distribution

Exercise 1: The time to failure (i hours) for a laser in a cytometry machine is modelled by an exponential distribution with lamda = 0.00004?

a. What is the probability that the laser will not fail in 20000 hours?

b. What is the probability that the laser will not last 30000 hours?

R code

```
> lamda = 0.00004
> 1-pexp(20000,lamda)
```

Probability that the laser will not fail in 20000 hours = 0.449329

Exponential Distribution

Exercise 2: The time between arrivals of taxis at busy intersection is exponentially distributed with a mean of 10 minutes. Simulate 50 time between arrivals of taxis to study the arrival pattern of taxis in a day?

R code

```
> mean = 10
> lamda = 1/mean
> iat = rexp(50,lamda)
> cbind(iat)
```

TEST
of
HYPOTHESIS

Hypothesis Testing Concepts Allow Us To ….

- **Properly handle uncertainty**
- **Minimize subjectivity**
- **Question assumptions**
- **Prevent the omission of important information**
- **Manage the risk of decision errors**

TEST OF HYPOTHESIS

- A hypothesis is a proposed explanation of a phenomenon or a commonly held belief.
- Hypothesis testing requires checking the validity of the explanation or the belief through data. Some examples of hypotheses are
  - Higher value invoices require longer payment time
  - Married women employees are likely to stay longer with the company than married male employees
  - Bidding frequently with lower average bid value is likely to lead to higher revenue growth compared to infrequent bidding with higher average bid value
  - Most customers who were given a retention offer would have stayed anyway

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 ($P_1 = P_2$)

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

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

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

β = Probability of making type II error

Power = $1 - \beta$: Probability of correctly rejecting a false null hypothesis

TEST OF HYPOTHESIS

1. Define the Practical Problem
2. State the Objectives (Create the Statistical Problem)
3. Establish the Hypotheses
   - State the Null Hypothesis (Ho)
   - State the Alternative Hypothesis (Ha).
4. Decide on appropriate statistical test (assumed probability distribution, z, t, or F).
5. State the α level (usually 5%), β level (usually 10-20%), effect size (δ) and establish the Sample Size
6. Develop the Sampling Plan, select samples, conduct test and collect data

12. Calculate the test statistic (z, t, or F) from the data.
13. Determine the probability of that calculated test statistic occurring by chance.
14. If that probability is less than α, reject Ho otherwise do not reject Ho.
15. Replicate results and translate statistical conclusion to practical solution.

TEST OF HYPOTHESIS

Test of Comparisons:

	Y = Continuous		Y = Discrete	
Comparison Type	**Mean**	**Variance**	**Defective**	**Defects**
Against Standard	1 Sample t	Chi-Square Test	1 sample p	1 sample defect rate
Between Two	2 Sample t OR Paired t	F-test	2 Sample p	2 sample defect rate
Among Many	ANOVA	Bartlett's Test	Chi-Square test	Chi-square

Note: The test mentioned for Y (Continuous) is applicable only when Y follows Normal Distribution. In case Y does not satisfy the Normality, then we need to use Non Parametric tests. For carrying out ANOVA, the condition of 'Equality of variance' to be satisfied.

Test of Modelling (X = Continuous):

Y = Continuous : Regression

Y = Discrete: Logistic Regression

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 $\mu = 5$ else $\mu \neq 5$

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 $\mu \neq 500$?

Conclusion:

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

TEST OF HYPOTHESIS

Methodology demo: To Test μ = 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 / $\sqrt{n}$)

If test statistic is close to 0, conclude that μ = 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

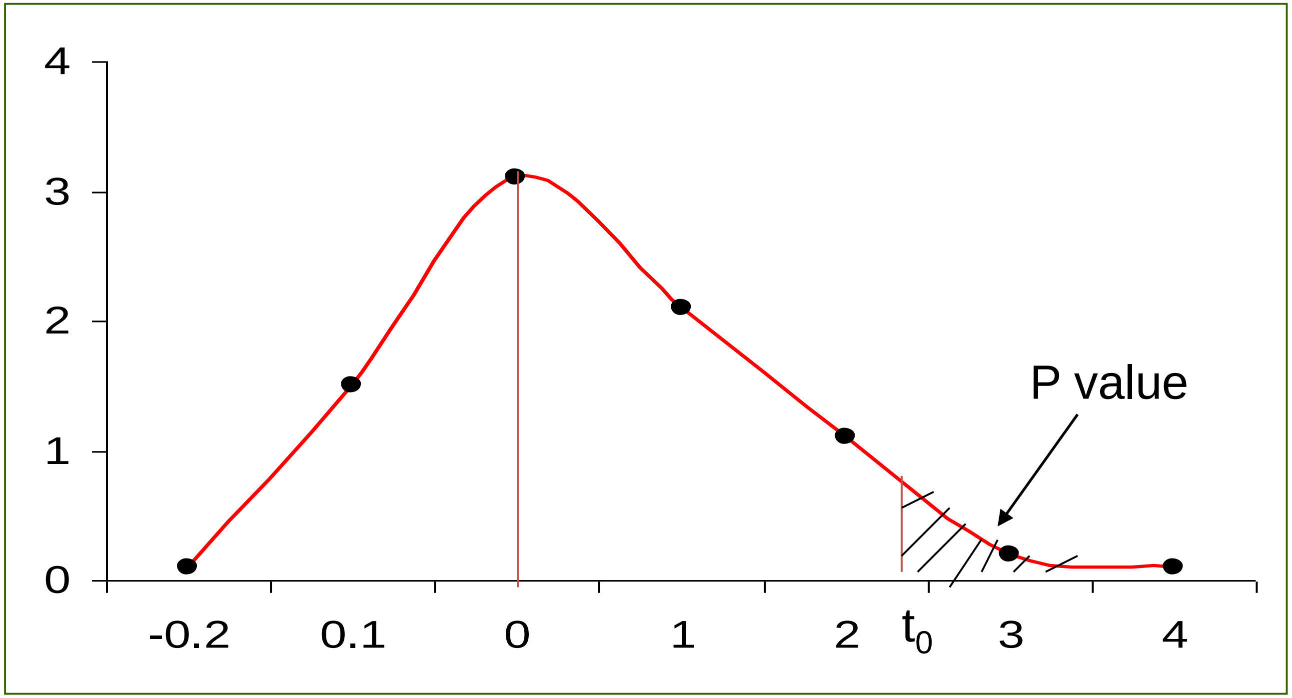
Methodology demo: To Test μ = Specified Value

P value

The probability that such evidence or result will occur when H0 is true

Based on the reference distribution of test statistic

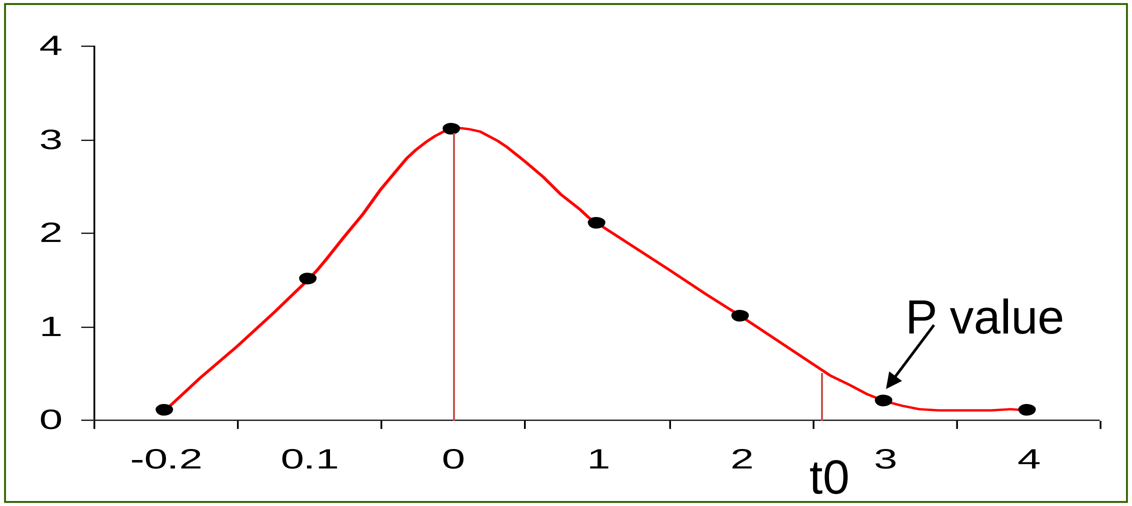
The tail area beyond the value of test statistic in reference distribution



TEST OF HYPOTHESIS

Methodology demo : To Test μ = Specified Value

P value



If test statistic t_0 is close to 0 then p will be high

If test statistic t_0 is not close to 0 then p will be small

If p is small , $p < 0.05$ (with $\alpha = 0.05$), conclude that $t \neq 0$, then

$\mu \neq$ Specified Value, H0 rejected

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: $\mu = 5$

H1: $\mu \neq 5$

Calculate xbar = 5.15

SD = 0.8515

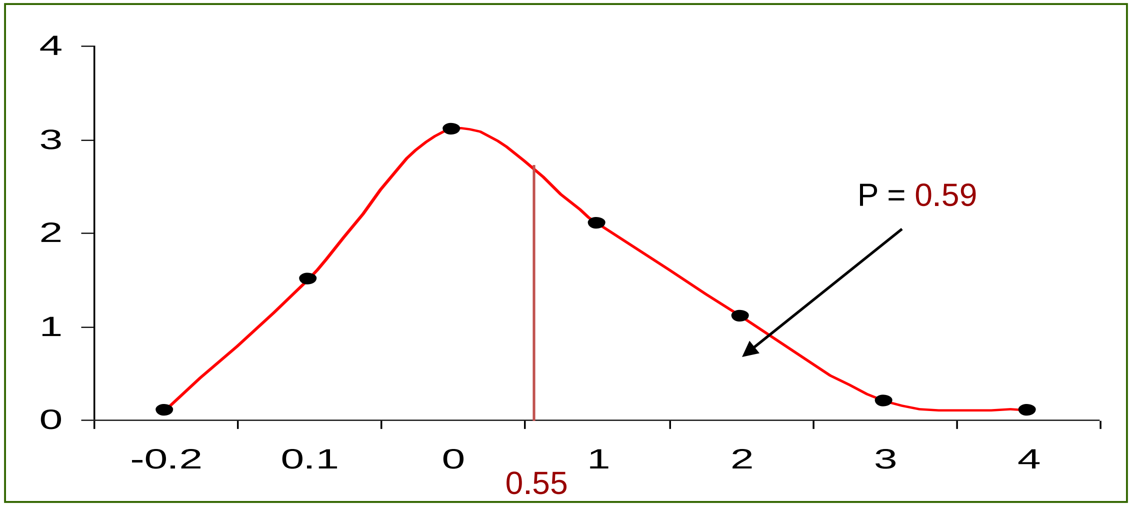
n = 10

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

TEST OF HYPOTHESIS

Example: To Test μ = Specified Value (μ = μ_0)

$t_0 = 0.5571$



P ≥ 0.05, hence μ = 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

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

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: $\text{Mean}_1 = \text{Mean}_2$ ($\mu_1 = \mu_2$)

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

or

H1: $\text{Mean}_1 > \text{Mean}_2$ ($\mu_1 > \mu_2$)

or

H1: $\text{Mean}_1 < \text{Mean}_2$ ($\mu_1 < \mu_2$)

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

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

Calculate p value from t distribution

If $p \geq 0.05$ then H0: $Mean_1 = Mean_2$ 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
```

Converting Outlet to Factor

```
> Outlet = factor(Outlet)
```

2 sample t Test

```
> t.test(Sales~Outlet, alternative = 'less')
```

Statistics	Value
t	0.9625
df	17.379
P value	0.8255

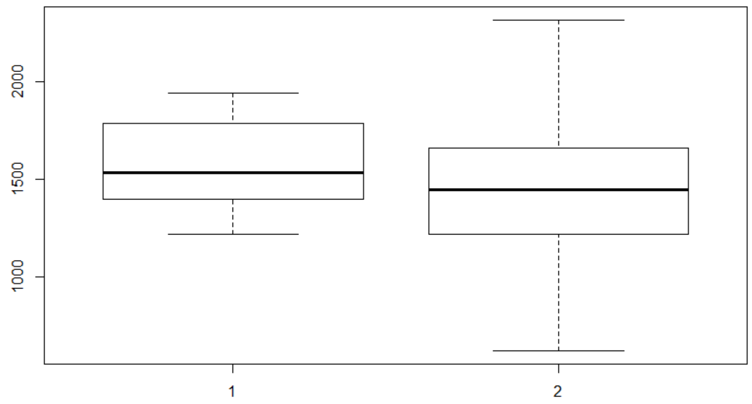
TEST OF HYPOTHESIS

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)
```



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

TEST OF HYPOTHESIS

Exercise 3: The data of 30 customers on credit card usage in INR1000, gender (1: male, 2: female) and whether they have done shopping or banking (1: yes , 2: no) with credit card are given in table below.

1. Check whether the average credit card usage is same for both gender?
2. Check whether the average credit card usage is same for those who do shopping with credit card and those who don't do shopping?
3. Check whether the average credit card usage is same for those who do banking with credit card and those who don't do banking?

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 Variance Test: Exercise 1

A super market chain has introduced a promotional activity in its selected outlets in the city to increase the sales volume. The outlets where promotional activity introduced are denoted by 1 and others by 2. Check for equality of variance?

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 Variance Test: Exercise 1

A super market chain has introduced a promotional activity in its selected outlets in the city to increase the sales volume. The outlets where promotional activity introduced are denoted by 1 and others by 2. Check for equality of variance?

Reading data to mydata

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

Converting Outlet to Factor

```
> Outlet = factor(Outlet)
```

2 Variance Test

```
> var.test(Sales~Outlet)
```

Statistics	Value
F	0.3196
Numerator df	11
Denominator df	11
P value	0.0713

TEST OF HYPOTHESIS

Two Variances test: Exercise 2

A bpo company have developed a new method for better utilization of its resources.. 10 observations on utilization from both methods is given below: Check whether both methods have same consistency with respect to utilization?

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

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 $t_0 = \dfrac{\bar{d}}{s_d / \sqrt{n}}$

Reject H0 if p – value < 0.05

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

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)
```

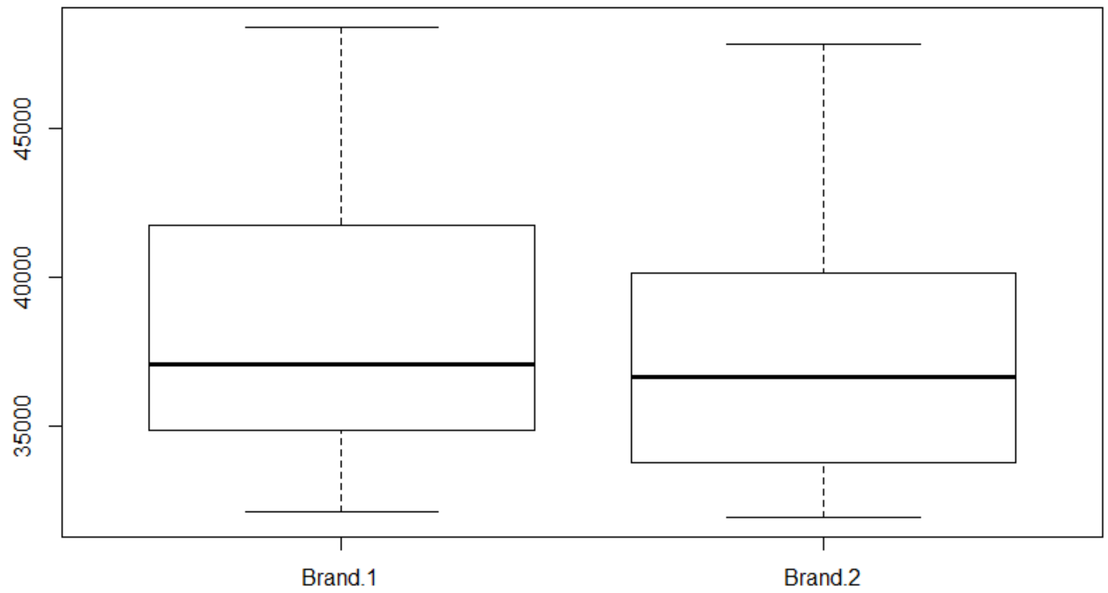
Statistics	Value
t	1.9039
df	7
P value	0.09863

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



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. One an average is the program successful?

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

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 ≠ 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 $\hat{p} = \text{specified value}(p_0)$

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

Calculate test statistic z by dividing $\hat{p} - \text{specified value}$ 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

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

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 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
India	6	551
Manila	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

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 - 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 $\neq$ 0

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

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

Reading the data and variable

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

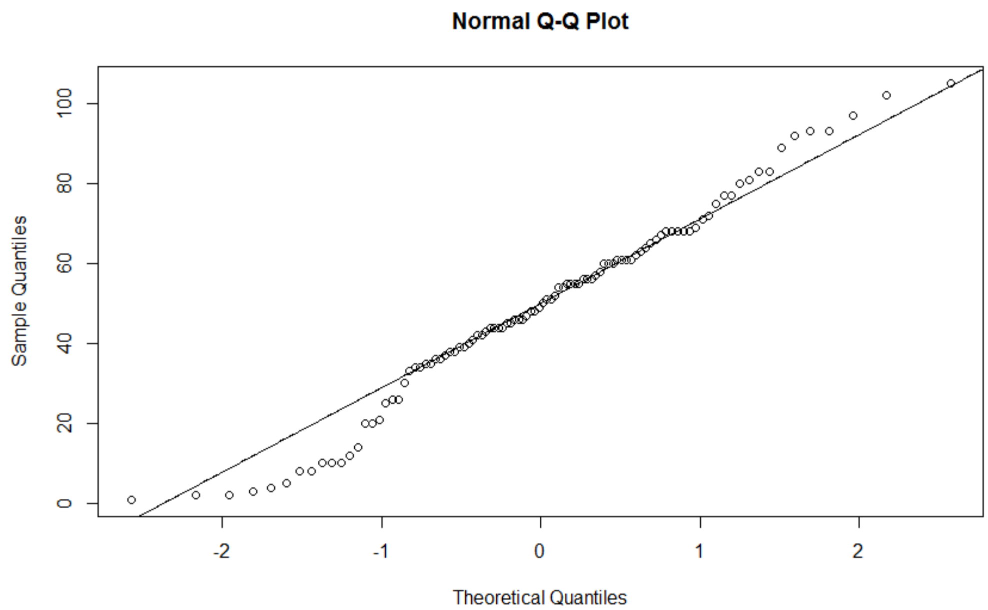
NORMALITY TEST

Normality test

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

Normality Check using Normal Q – Q plot

```
> qqnorm(PT)
> qqline(PT)
```



NORMALITY TEST

Normality test

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

Normality Check using Shapiro – Wilk test

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

Statistics	Value
W	0.9804
p value	0.1418

NORMALITY TEST

Normality test

Exercise 2 : The time taken to respond to customer complaints 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

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?

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 location of shelf on sales revenue

ANALYSIS OF VARIANCE

One Way ANOVA : Example

An electronics and home appliance chain suspect the location of shelves where television sets are kept will influence the sales revenue. The data on sales revenue in lakhs from the television sets when they are kept at different locations inside the store are given in sales revenue data file. The location is denoted as 1:front, 2: middle & 3: rear. Verify the doubt? The data is given in Sales_Revenue_Anova.csv.

Factor: Location(A)

Levels : front, middle, rear

Response: Sales revenuec

ANALYSIS OF VARIANCE

One Way ANOVA : Example

Step 1: Calculate the sum, average and number of response values for each level of the factor (location).

Level 1 Sum(A_1):

Sum of all response values when location is at level 1 (front)

$$= 1.55 + 2.36 + 1.84 + 1.72 = 7.47$$

nA_1: Number of response values with location is at level 1 (front) = 4

Average: Sum of all response values when location is at level 1 / number of response values with location is at level 1

$$= A_1 / nA_1 = 7.47 / 4 = 1.87$$

	Level 1 (front)	Level 2 (middle)	Level 3 (rear)
Sum	A_1: 7.47	A_2: 30.31	A_3: 15.55
Number	nA_1: 4	nA_2: 8	nA_3: 6
Average	1.87	3.79	2.59

ANALYSIS OF VARIANCE

One Way ANOVA : Example

Step 2: Calculate the grand total (T)

T = Sum of all the response values

= 1.55 + 2.36 + - - - + 2.72 + 2.07 = 53.33

Step 3: Calculate the total number of response values (N)

N = 18

Step 4: Calculate the Correction Factor (CF)

CF = $(\text{Grand Total})^2$ / Number of Response values

$= T^2 / N = (537.33)^2 / 18 = 158.0049$

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

TSS = Sum of square of all the response values - CF

$= 1.55^2 + 2.36^2 + - - - + 2.72^2 + 2.07^2 – 158.0049$

= 15.2182

ANALYSIS OF VARIANCE

One Way ANOVA : Example

Step 6: Calculate the between (factor) sum of square

$$SS_A = A_1^2 / nA_1 + A_2^2 / nA_2 + A_3^2 / nA_3 - CF$$

$$= 7.47^2 / 4 + 30.31^2 / 8 + 15.55^2 / 4 – 158.0049 = 11.0827$$

Step 7: Calculate the within (error) sum of square

SS_e = Total sum of square – between sum of square

$$= TSS - SS_A = 15.2182 – 11.0827 = 4.1354$$

Step 8: Calculate degrees of freedom (df)

Total df = Total Number of response values – 1 = 18 - 1 = 17

Between df = Number of levels of the factor - 1 = 3 - 1 = 2

Within df = Total df – Between df = 17 - 2 = 15

ANALYSIS OF VARIANCE

One Way ANOVA : Example – ANOVA Table

Source	df	SS	MS	F	F Crit	P value
Between	2	11.08272	5.541358	20.09949	3.68	0.0000
Within	15	4.135446	0.275696			
Total	17	15.21816				

MS = SS / df : F = $MS_{Between}$/ MS_{Within}

F Crit =finv (probability, between df, within df) , probability = 0.05

P value = fdist (F, between df, within df)

One Way ANOVA : **Decision Rule**

If p value < 0.05, then the factor has significant effect on the process output or response. In this example as p value is < 0.05 means location has significant effect on sales revenue

Meaning: When the factor is changed from one level to another level, there will be significant change in the mean response. Here the sales revenue is not same for different locations like front, middle & rear.

ANALYSIS OF VARIANCE

One Way ANOVA : R Code

Reading data and variables to R

```
> mydata = Sales_Revenue_Anova
> location = mydata$Location
> revenue = mydata$Sales.Revenue
```

Converting location to factor

```
> location = factor(location)
```

Computing ANOVA table

```
> fit = aov(Revenue ~ location)
> summary(fit)
```

ANALYSIS OF VARIANCE

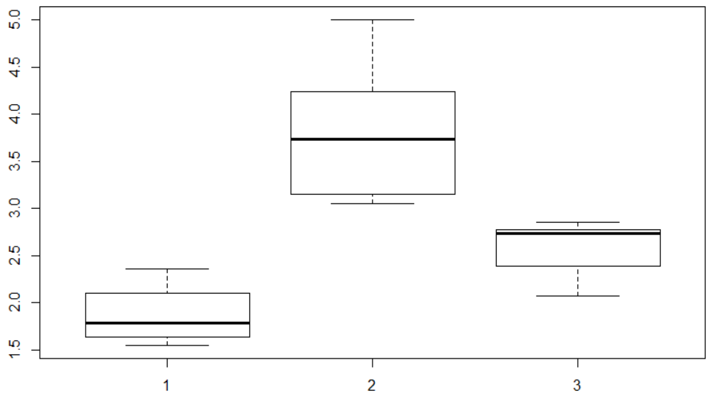
One Way ANOVA : Example Result

The expected sales revenue for different location under study is equal to level averages.

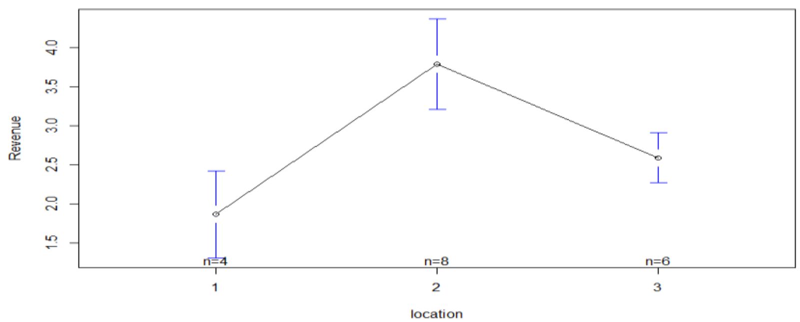
> aggregate(Revenue ~ location, FUN = mean)

Location	Expected Sales Revenue
Front	1.8675
Middle	3.78875
rear	2.591667

> boxplot(Revenue ~ location)



> library(gplots)
> plotmeans(Revenue ~ location)



ANALYSIS OF VARIANCE

ANOVA logic:

Two Types of Variations:

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

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.

	Location		
	Front	Middle	rear
Sales Revenue	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 9.80	2.30 1.91 1.40 1.48

ANALYSIS OF VARIANCE

ANOVA logic :

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

Measure of variation between levels: MS of the factor ($MS_{between}$)

Measure of variation within levels: MS Error (MS_{within})

To check whether a factor is significant:

Compare MS of between with MS within

i.e. Calculate $F = MS_{between} / MS_{within}$

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(Revenue, l ocation, data = mydata)
```

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

Bartlett's K^2 Statistic	df	p value
3.8325	2	0.1472

Since p value = 0.1472 > 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 turning 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					

Example: A branded apparel manufacturing company has collected the data from 50 customers on usage, gender, awareness of brand and preference of the brand. Usage has been coded as 1, 2 ,and 3 representing light, medium and heavy usage. The gender has been coded as 1 for female and 2 for male users. The attitude and preference are measured on a 7 point scale (1: unfavorable to 7 : very favorable). The data is given in apparel_data.csv file .

1. Does male and female differ in their usage?
2. Does male and female differ in their awareness of the brand?
3. Does male and female differ in their preference?
4. Does higher the awareness means higher preference?
5. Does high awareness and high preference leads to heavy usage?

a. Reading the file and converting variables to factors

```
> mydata = Apparel_Data
> usage = factor(mydata$Usage)
> gender = factor(mydata$Gender)
> awareness = factor(mydata$Awareness)
> preference = factor(mydata$Preference)
```

b. Constructing cross tabulation of Gender vs. Usage

```
> mytable = table(usage, gender)
> print(mytable)
```

Or

```
> library(gmodels)
> CrossTable(gender, usage, prop.r = FALSE, prop.c = FALSE, prop.t = FALSE,
  prop.chisq=FALSE)
```

Gender	Usage			
	1	2	3	Total
1	15	6	5	26
2	6	6	12	24
Total	21	12	17	50

c. Constructing cross tabulation of Gender vs. Usage – cell proportions

```
> mytable = table(usage, gender)
> prop.table(mytable)
```

Gender	Usage			
	1	2	3	Total
1	0.30	0.12	0.10	0.52
2	0.12	0.12	0.24	0.48
Total	0.42	0.24	0.34	1.00

d. Constructing cross tabulation of Gender vs. Usage – row proportions

```
> mytable = table(usage, gender)
> prop.table(mytable, 1)
```

Gender	Usage			
	1	2	3	Total
1	0.58	0.23	0.19	1.00
2	0.25	0.25	0.50	1.00

d. Constructing cross tabulation of Gender vs. Usage – column proportions

```
> mytable = table(usage, gender)
> prop.table(mytable, 2)
```

Gender	Usage		
	1	2	3
1	0.72	0.50	0.29
2	0.28	0.50	0.71
Total	1.00	1.00	1.00

5. Constructing three way cross tabulation of Awareness, Preference and Usage

```
> mytable = table(awareness, preference, usage)
> ftable(mytable)
```

Exercise 1: An ITeS company has collected following information from its customers through survey. The data has been collected in 5 point scale (1: Very dissatisfied to 5: Very satisfied). The survey questions are given below and data is given in Csat_data file. Check whether the questions 1 to 9 are related to overall satisfaction

1. Team’s ability to meet service level agreements
2. Team’s ability to deliver seamlessly in the event of changes (volume fluctuations, resource movement etc)
3. Team’s operational performance
4. Team's application of process knowledge
5. Team's communication with customer
6. Team’s effectiveness in handling escalations
7. Team’s flexibility and responsiveness to special service requests
8. Team’s contribution to customer’s business requirements
9. Effectiveness of the reviews around operations delivery
10. Overall with team’s service

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

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

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

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 $\chi^2 = 161.98$

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

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: A branded apparel manufacturing company has collected the data from 50 customers on usage, gender, awareness of brand and preference of the brand. Usage has been coded as 1, 2 ,and 3 representing light, medium and heavy usage. The gender has been coded as 1 for female and 2 for male users. The attitude and preference are measured on a 7 point scale (1: unfavorable to 7 : very favorable). The data is given in apparel_data.csv file .

1. Estimate the relation between gender and usage?
2. Estimate the relation between gender and awareness of the brand?
3. Estimate the relation between gender and preference?
4. Does higher the awareness means higher preference?

a. Reading the file and converting variables to factors

```
> mydata = Apparel_Data
> usage = factor(mydata$Usage)
> gender = factor(mydata$Gender)
> awareness = factor(mydata$Awareness)
> preference = factor(mydata$Preference)
```

b. Constructing cross tabulation of Gender vs. Usage

```
> mytable = table(usage, gender)
> print(mytable)
```

Gender	Usage			
	1	2	3	Total
1	15	6	5	26
2	6	6	12	24
Total	21	12	17	50

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

```
> chisq.test(mytable)
```

Statistics	Value
Chi Square	6.6702
df	2
P value	0.03561

Fisher's Exact test

When one or more of expected frequencies are less than 5

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

```
> fisher.test(mytable)
```

Statistics	Value
P value	0.0348

e. Measures of Association - Gender vs. Usage

```
> library(vcd)
> assocstats(mytable)
> kappa(mytable)
```

	Chi Square	df	p - value
Likelihood Ratio	6.8747	2	0.032149
Pearson	6.6702	2	0.035612

Statistics	Value
Phi-Coefficient	0.365
Contingency Coefficient	0.343
Cramer's V	0.365
kappa	

CHI SQUARE TEST

Exercise 1: An ITeS company has collected following information from its customers through survey. The data has been collected in 5 point scale (1: Very dissatisfied to 5: Very satisfied). The survey questions are given below and data is given in Csat_data file. Check whether the questions 1 to 9 are related to overall satisfaction?

1. Team's ability to meet service level agreements
2. Team's ability to deliver seamlessly in the event of changes (volume fluctuations, resource movement etc)
3. Team's operational performance
4. Team's application of process knowledge
5. Team's communication with customer
6. Team's effectiveness in handling escalations
7. Team's flexibility and responsiveness to special service requests
8. Team's contribution to custome'rs business requirements
9. Effectiveness of the reviews around operations delivery
10. Overall satisfaction with team's service

PREDICTIVE ANALYTICS

PREDICTIVE ANALYTICS

Methods

1. Parametric Methods
2. Non parametric Methods

Parametric Methods

Independent Variables (Xs)	Dependant Variables (Y)	Techniques
Continuous	Continuous	Multiple Linear Regression
Discrete	Continuous	Dummy Variable Regression
Continuous	Discrete	Logistic Regression

CORRELATION
&
REGRESSION

CORRELATION & REGRESSION

Correlation:

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

Type and degree of relationship between two variables.

Correlation: Usage

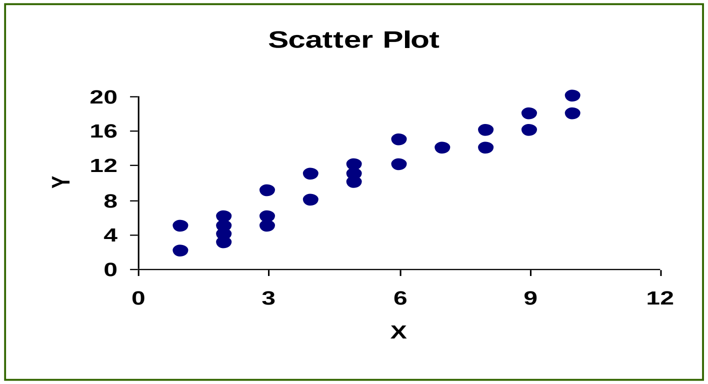
Explore the relationship between the output characteristic and input or process variable.

Output variable : Y : Dependent variable

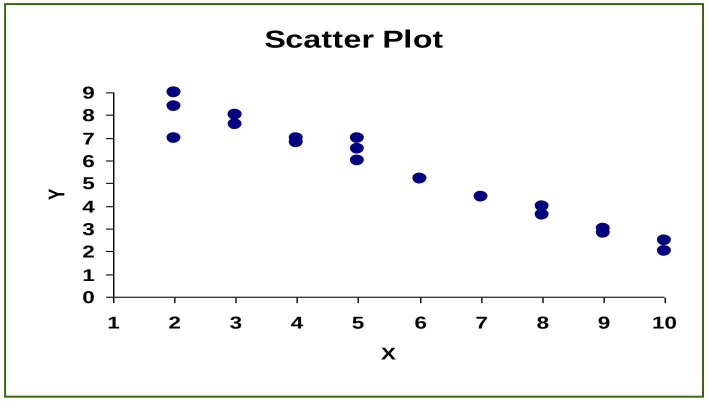
Input / Process variable : X : Independent variable

CORRELATION & REGRESSION

Positive Correlation: Y increases as X increases & vice versa

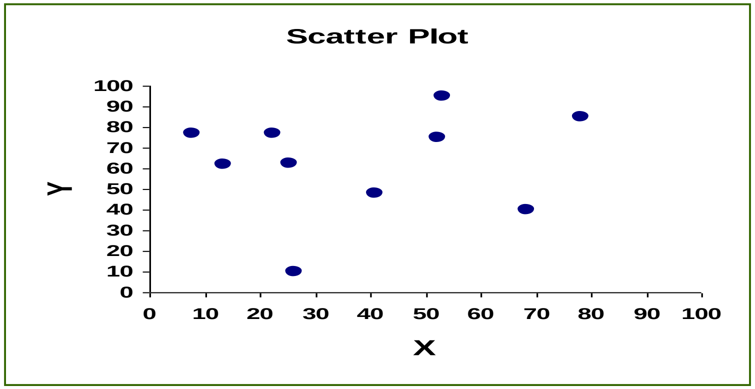


Negative Correlation: Y decreases as X increases & vice versa

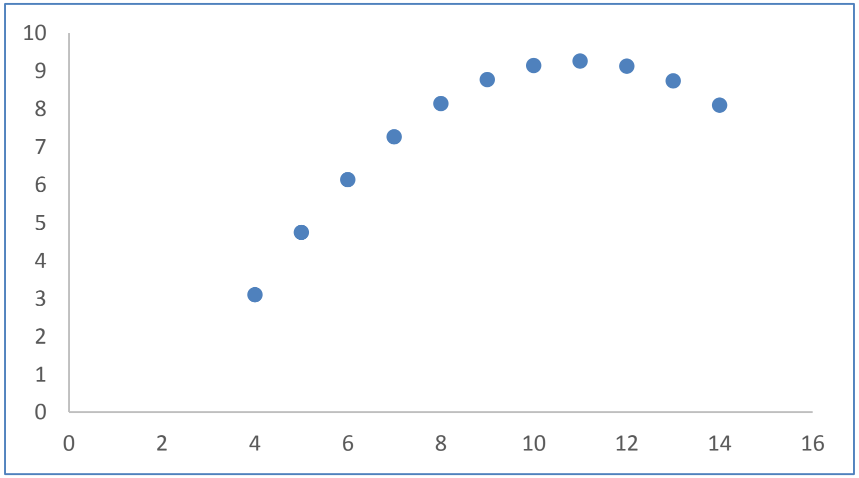


CORRELATION & REGRESSION

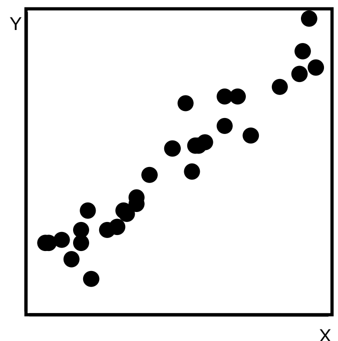
No Correlation: Random Distribution of points



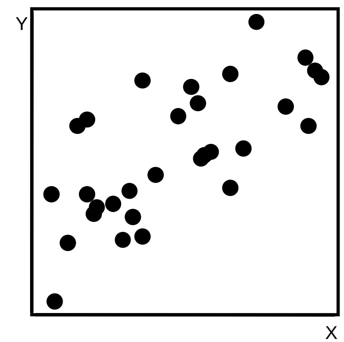
Non Linear Correlation: Curvature form of points



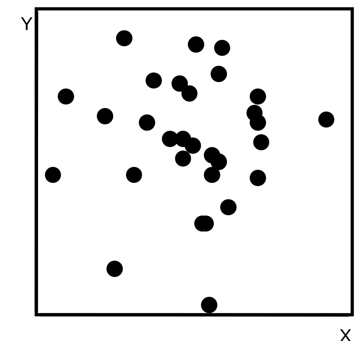
Is there any correlation ?



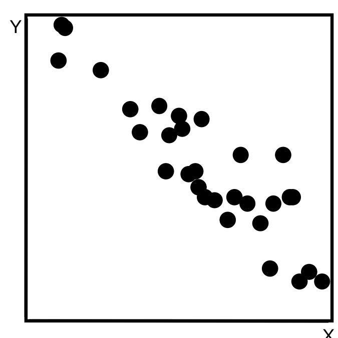
Plot - A



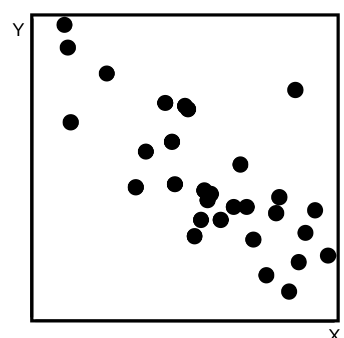
Plot - B



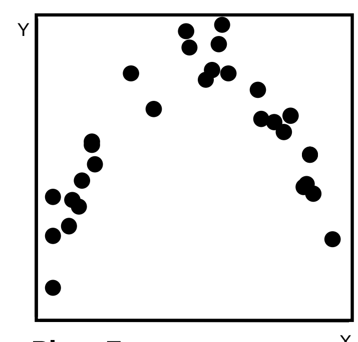
Plot - C



Plot - D



Plot - E



Plot - F

CORRELATION & REGRESSION

Measure of Correlation: Coefficient of Correlation

Symbol : r

Range : -1 to 1

Sign : Type of correlation

Value : Degree of correlation

$$r = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{\sqrt{\sum (X - \overline{X})^2} \sqrt{\sum (Y - \overline{Y})^2}}$$

CORRELATION & REGRESSION

Coefficient of Correlation Computation :

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

Coefficient of Correlation Computation :

SL No.	x – Mean x	y – Mean y	Product	$(x - \text{Mean } x)^2$	$(y - \text{Mean } y)^2$
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 / √Sxx.Syy = -54 / √(28 x 104.83) = -0.9967

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

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?

R-Code:

Reading the data and variables

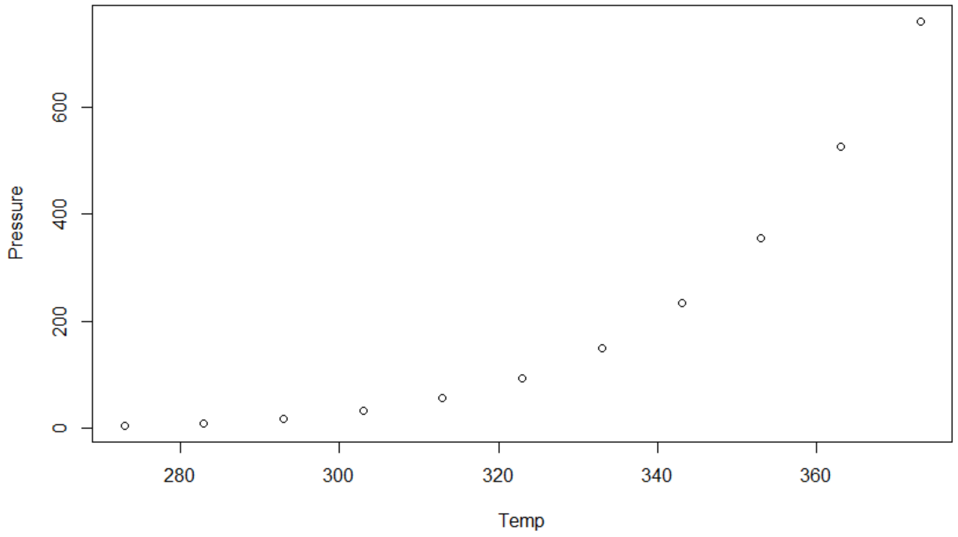
```
> mydata = Correlation
> Temp = mydata$Temperature
> Pressure = mydata$Vapor.Pressure
```

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)



Computing correlation coefficient

> cor(Temp, Pressure)

Statistics	Value
r	0.893

CORRELATION & REGRESSION

Regression

Correlation helps

To check whether two variables are related

If related

Identify the type & degree of relationship

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

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

$$\mathbf{y = 1 + 3x}$$

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

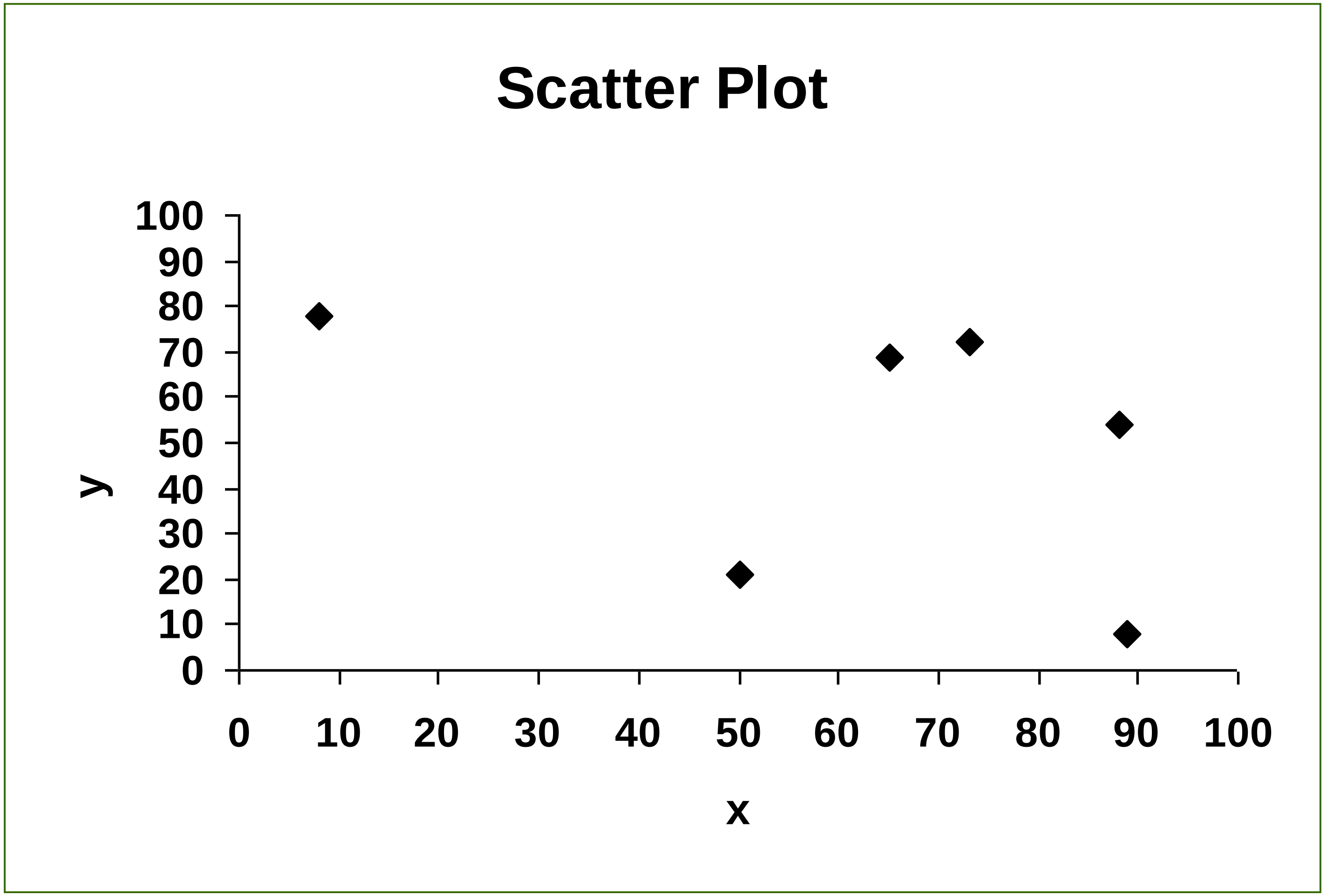
If p value < 0.05, then H0 is rejected & y can be modeled with x

Regression illustration: Example

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

CORRELATION & REGRESSION

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



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

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

CORRELATION & REGRESSION

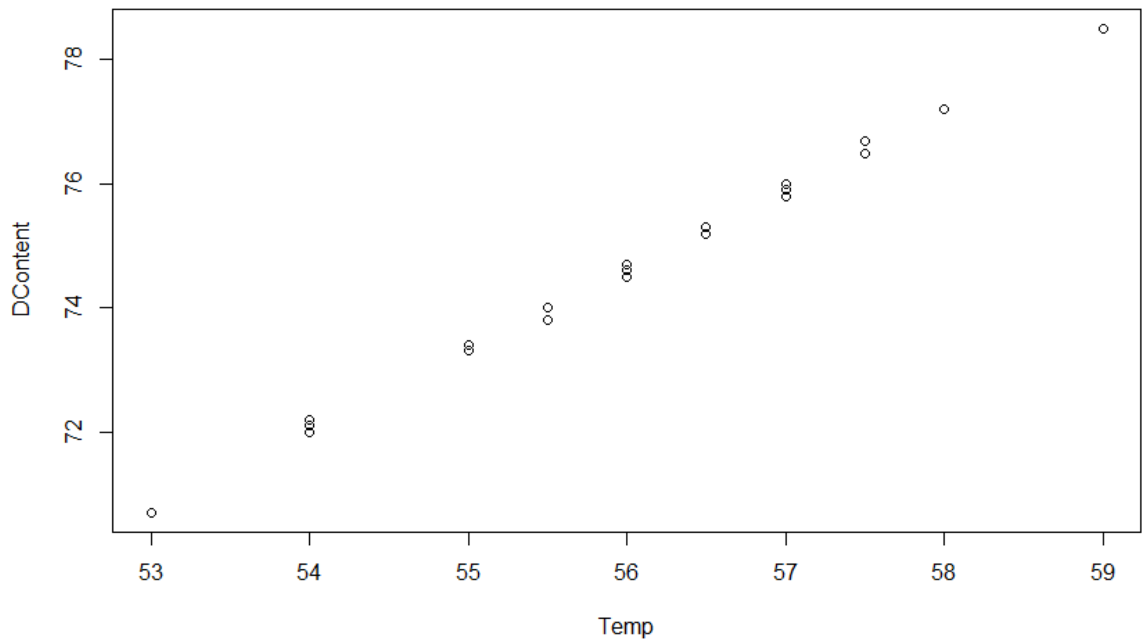
Exercise 1: The data from the pulp drying process is given in the file DC_Simple_Reg.csv. The file contains data on the dry content achieved at different dryer temperature. Develop a prediction model for dry content in terms of dryer temperature.

1. Reading the data and variables

```
> mydata = DC_Simple_Reg
> Temp = mydata$Dryer.Temperature
> DContent = mydata$Dry.Content
```

2. Constructing Scatter Plot

```
> plot(Temp, DContent)
```



3. Computing Correlation Matrix

```
> cor(Temp, DContent)
```

Attribute	Dry Content
Temperature	0.9992

Remark:

Correlation between y & x need to be high (preferably 0.8 to 1 to -0.8 to -1.0)

4: Performing Regression

```
> model = lm(DContent ~ Temp)
> summary(model)
```

Statistic	Value	Criteria
Residual standard error	0.07059	
Multiple R-squared	0.9984	> 0.6
Adjusted R-squared	0.9983	> 0.6

Model	df	F	p value
Regression	1	24497	0.000
Residual	40		
Total	41		

Criteria:
P value < 0.05

CORRELATION & REGRESSION

4: Performing Regression

Attribute	Coefficient	Std. Error	t Statistic	p value
Intercept	2.183813	0.463589	4.711	0.00
Temperature	1.293432	0.008264	156.518	0.00

Interpretation

The p value for independent variable need to be < significance level α (generally $\alpha = 0.05$)

Model: Dry Content = 2.183813 + 1.293432 x Temperature

5: Regression ANOVA

```
> anova(model)
```

ANOVA

Source	SS	df	MS	F	p value
Temp	122.057	1	122.057	24497	0.000
Residual	0.199	40	0.005		
Total	122.256	41			

Criteria: P value < 0.05

5: Residual Analysis

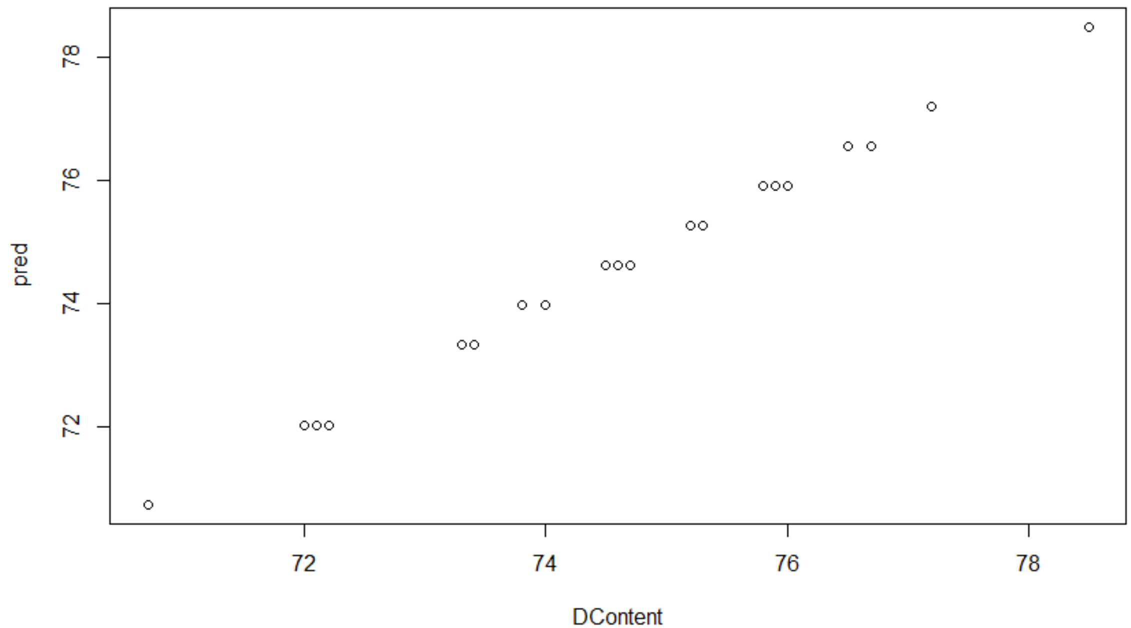
```
> pred = fitted(model)
> Res = residuals(model)
> write.csv(pred,"D:/Infosys/DataSets/Pred.csv")
> write.csv(Res,"D:/Infosys/DataSets/Res.csv")
```

SL No.	Fitted	Residuals	SL No.	Fitted	Residuals
1	73.32259	-0.02259	22	74.61602	-0.01602
2	74.61602	-0.01602	23	75.26274	-0.06274
3	73.96931	0.030693	24	73.96931	0.030693
4	78.49632	0.00368	25	75.90946	-0.00946
5	74.61602	-0.01602	26	75.26274	0.03726
6	73.96931	0.030693	27	73.96931	0.030693
7	75.26274	-0.06274	28	78.49632	0.00368
8	77.20289	-0.00289	29	76.55617	-0.05617
9	75.90946	-0.00946	30	74.61602	-0.11602
10	74.61602	-0.01602	31	75.90946	0.090544
11	73.32259	-0.02259	32	76.55617	-0.05617
12	75.90946	-0.00946	33	76.55617	0.143828
13	75.90946	0.090544	34	75.90946	0.090544
14	74.61602	-0.01602	35	75.90946	-0.10946
15	74.61602	0.083977	36	73.96931	-0.16931
16	74.61602	-0.11602	37	73.32259	-0.02259
17	70.73573	-0.03573	38	74.61602	-0.01602
18	72.02916	-0.02916	39	73.32259	0.077409
19	72.02916	0.070841	40	75.90946	0.090544
20	72.02916	0.170841	41	73.96931	0.030693
21	70.73573	-0.03573	42	75.26274	-0.06274

5: Residual Analysis

Scatter Plot: Actual Vs Predicted (fit)

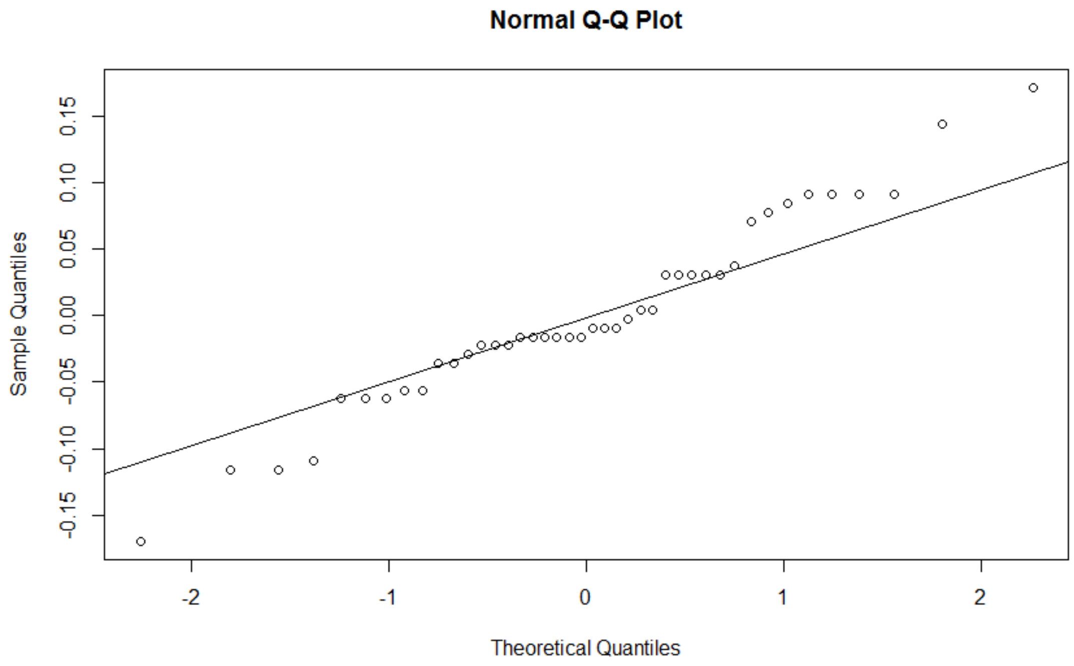
```
> plot(DContent, pred)
```



5: Residual Analysis

Normality Check on residuals

```
> qqnorm(Res)
> qqline(Res)
```



Residuals should be normally distributed or bell shaped

5: Residual Analysis

Normality Check on residuals

```
> shapiro.test(Res)
```

Shapiro-Wilk normality Test:	
W	p value
0.9693	0.3132

Residuals should be normally distributed or bell shaped

CORRELATION & REGRESSION

5: Residual Analysis

```
> plot(pred, Res)
> plot(Temp, Res)
```

Residuals should be independent and stable

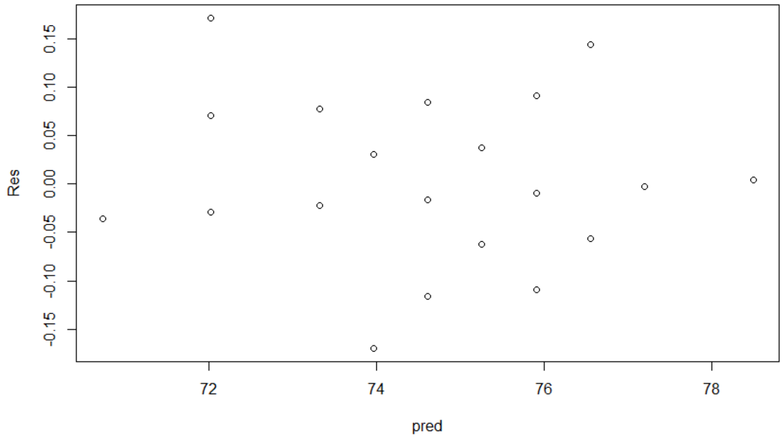
Plot the residuals against fitted value. The points in the graph should be scattered randomly and should not show any trend or pattern. The residuals should not depend in anyway on the fitted value.

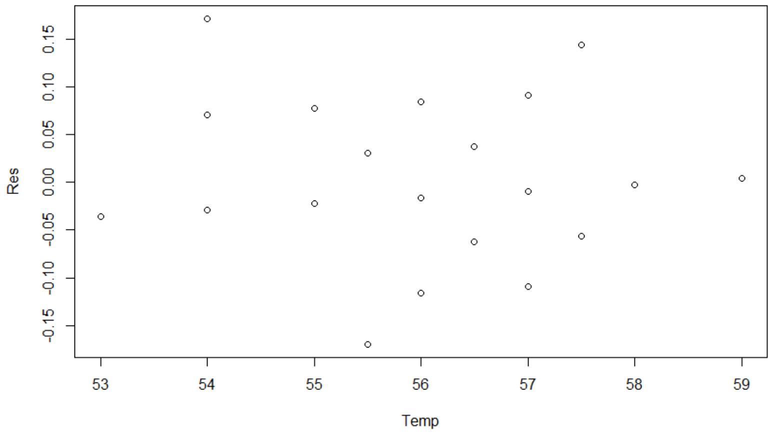
If there is a pattern then a transformation such as log y or $\sqrt{y}$ to be used

Similarly the residuals shall not depend on x. This can be checked by plotting residuals vs x. A pattern in this plot is an indication that the residuals are not independent of x.

CORRELATION & REGRESSION

Residual Analysis





There is no trend or pattern on residuals vs fitted value ,residuals vs observation order or residuals vs x plot. Hence the assumptions of independence and stability of residuals are satisfied.

6: Outlier test

Observations with Bonferonni p – value < 0.05 are potential outliers

```
> library(car)
> outlierTest(model)
```

Observation	Studentized Residual	Bonferonni p value
20	2.723093	0.40417

7: Leave One Out Cross Validation (LOOCV)

- Split the data into two parts : training data and test data

  Test data consists of only one observation (x_1, y_1)

  Training data consists of the remaining n – 1 observations namely (x_2, y_2) , (x_3, y_3) , - - -, (x_n, y_n)
- Develop the model using n – 1 training data observations and predict the response y_1 of the test data observation

  Compute the residuals and mean square error $MSE_1 = (y_{1actual} – y_{1pred})^2$
- Repeat the process by taking (x_1, y_1) as test data and the remaining n – 1 observations as training data
- Compute MSE_2
- Repeating the procedure n times produces n squared errors MSE_1, MSE_2, - - -, MSE_n
- LOOCV estimate of the test MSE is the average of these n test error estimates

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} MSE_i$$

7: Leave One Out Cross Validation (LOOCV)

```
> library(boot)

> attach(mydata)

> mymodel = glm(Dry.Content ~ Dryer.Temperature)

> valid = cv.glm(mydata, mymodel)

> valid$delta[1]
```

Statistic	Value
Delta	0.005201004

CORRELATION & REGRESSION

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)

CORRELATION & REGRESSION

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

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

CORRELATION & REGRESSION

Step 2: Regression Output

Statistic	Value	Criteria
Adjusted R Square	0.7766	≥ 0.6

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

Step 2: Regression Output

ANOVA

Source	SS	df	MS	F	p value
Time	6777.8	1	6777.8	53.9872	0.000
Temp	19.3	1	19.3	0.1534	0.702
Residual	1632.1	13	125.5		

Criteria: P value < 0.05

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

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

CORRELATION & REGRESSION

Step 3: Residual Analysis

SL No.	Temperature	% Yield	Predicted	Time
1	190	35.0	36.22	130
2	176	81.7	76.10	174
3	205	42.5	39.84	134
4	210	98.3	91.51	191
5	230	52.7	67.94	165
6	192	82.0	94.23	194
7	220	34.5	48.00	143
8	235	95.4	86.98	186
9	240	56.7	44.38	139
10	230	84.4	88.79	188
11	200	94.3	77.01	175
12	218	44.3	59.79	156
13	220	83.3	90.61	190
14	210	91.4	79.73	178
15	208	43.5	38.03	132
16	225	51.7	52.53	148

Step 3: Residual Analysis:

Shapiro-Wilk normality Test: Yield data	
W	p value
0.9449	0.4132

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

7: Leave One Out Cross Validation (LOOCV)

```
> library(boot)

> attach(mydata)

> mymodel = glm(X.Yield ~ Time)

> myvalidation = cv.glm(mydata, mymodel)

> myvalidation$delta[1]
```

Statistic	Value
Delta	128.8541

CORRELATION & REGRESSION

Exercise : 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?

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

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

CORRELATION & REGRESSION

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			

CORRELATION & REGRESSION

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)

```
> vif(mymodel)
```

x	VIF
Temperature	12.23
Time	12.33
Kappa #	1.062

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)

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

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

REGRESSION ANALYSIS

Tackling Multi collinearity:

Method 1: Stepwise Regression

```
> pred = predict(mymodel)
> res = residuals(mymodel)
 > cbind(X..Conversion, pred, res)
> mse = mean(res^2)
> rmse = sqrt(mse)
```

Statistic	Value
Mean Square Error (MSE)	10.7
Root Mean Square Error (RMSE)	3.27

k fold Cross Validation

Steps

1. Divide the data set into k equal subsets
2. Keep one subset (sample) for model validation
3. Develop the model using all the other k – 1 subsets data put together
4. Predict the responses for the test data and compute residuals
5. Return the test sample back to the original data set and take another subset for model validation
6. Go to step 3 and continue until all the subsets are tested with different models
7. Compute the overall Root Mean Square Residuals. RMSE of validation should not be high compared to the original model developed with all the data points together.

Note: when k = n, then k fold cross validation is same as leave one out cross validation

k fold Cross Validation

R code

```
> library(DAAG)
> cv.lm(mymodel, m = 16)
> cv.lm(mymodel, df = mydata, m = 16)
```

m: number of validations required. M = 16 = n, hence equal to leave one out cross validation

Model	MSE	RMSE
Original	10.7	3.27
Cross Validation	19.6	4.43

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 to measure the effect of gender and income on attitude towards vocation. Data was collected from 30 respondents and is given in Travel_dummy_reg file. Attitude towards vocation is measured on a 9 point scale. Gender is coded as male = 1 and female = 2. Income is coded as low=1, medium = 2 and high = 3. Develop a model for attitude towards vocation in terms of gender and Income?

Regression with dummy variables

Variable		Dummy
Gender	Code	gender_Code
Male	1	0
Female	2	1

Variable		Dummy	
Income	Code	Income1	Income 2
Low	1	0	0
Medium	2	1	0
High	3	0	1

Regression with dummy variables

Read the fie and variables

```
> mydata =  Travel_dummy_Reg
> mydata = mydata[,2:4]
> gender = mydata$Gender
> Income = mydata$Income
> Attitude = mydata$Attitude
```

Converting categorical x's to factors

```
> gender = factor(gender)
> income = factor(income)
```

Regression with dummy variables – Output

- mymodel = lm(attitude ~ genser + income)
- summary (mumodel)

Multiple R^2	0.8603
Adjusted R^2	0.8442
F Statistics	53.37
P value	0.00

	Estimate	Std. Error	t value	p value
(Intercept)	2.4	0.3359	7.145	0.00000
gender2	-1.6	0.3359	-4.763	0.00006
income2	2.8	0.4114	6.806	0.00000
income3	4.8	0.4114	11.668	0.00000

> anova (mumodel)

	Df	Sum Sq	Mean Sq	F	p value
gender	1	19.2	19.2	22.691	0.0001
income	2	116.27	58.133	68.703	0.0000
Residuals	26	22	0.846		

MODELING NONLINEAR RELATIONS

MODELING NONLINEARRELATIONS

The linear regression is fast and powerful tool to model complex phenomena

But makes several assumptions about the data including the assumption of linear relationship exists between predictors and response variable.

When these assumptions are violated, the model breaks down quickly

MODELING NONLINEAR RELATIONS

The linear model y = xβ + ε is general model

Can be used to fit any relationship that is linear in the unknown parameter β

Examples:

$$y = \beta_0 + \beta_1 x_1 + --- + \beta_k x_k + \varepsilon$$

$$y = \beta_0 + \beta_1 x_1 + \beta_{11} x_1^2 + \varepsilon$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \varepsilon$$

In general

$$y = \beta_0 + \beta_1 f(x) + \varepsilon$$

where f(x) can be 1/x, √x, log(x), e^x , etc

MODELING NONLINEAR RELATIONS

Detection of non linear relation between predictor x and response variable y

Scatter Plot:

The plotted points are not lying lie in a straight line is an indication of non linear relationship between predictor and dependant variable

Component Residual Plots:

An extension of partial residual plots

Partial residual plots are the plots of residuals of one predictor against dependant variable

Component residual plots(crplots) adds a line indicating where the best fit line lies.

A significant difference between the residual line and the component line indicate that the predictor does not have a linear relationship wit the dependent variable

MODELING NONLINEAR RELATIONS

Example : The data given in Nonlinear_Thrust.csv represent the trust of a jet – turbine engine (y) and 3 predictor variables: x_3 = fuel flow rate, x_4 = pressure, and x_5 = exhaust temperature. Develop a suitable model for thrust in terms of the predictor variables.

Read Data

```
> attach(mydata)
> cor(mydata)
```

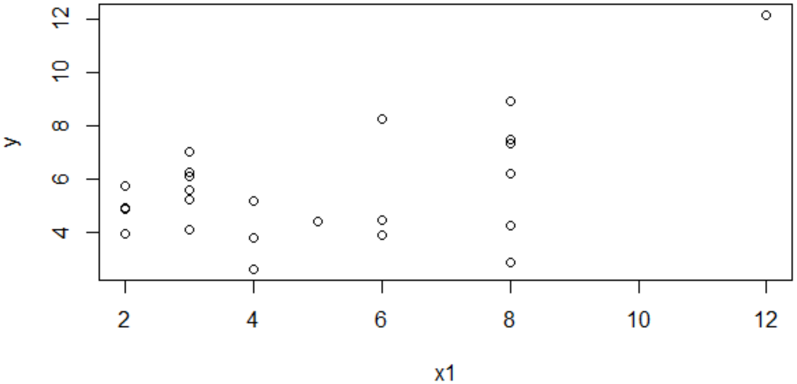
	x1	x2	x3	y
x1	1.00	0.40	-0.20	0.54
x2	0.40	1.00	-0.30	-0.36
x3	-0.20	-0.30	1.00	0.35
y	0.54	-0.36	0.35	1.00

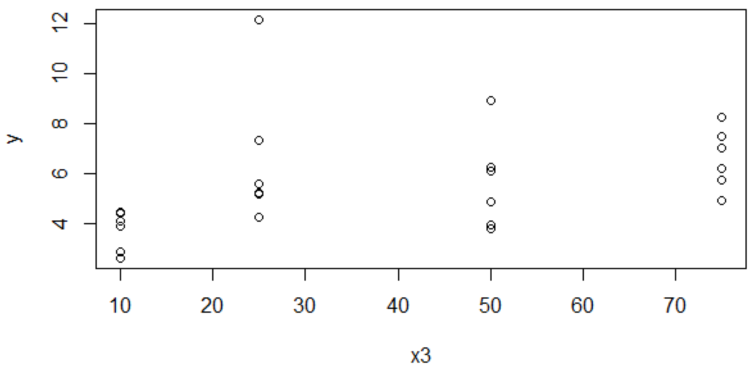
There is no strong correlation between y and x's

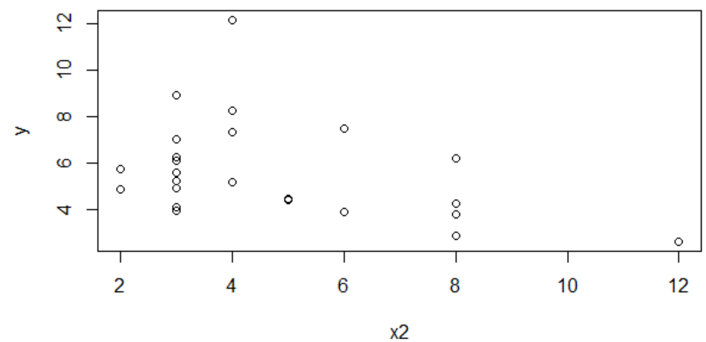
MODELING NONLINEAR RELATIONS

Draw Scatter plots

> plot(x1,y)
> plot(x2,y)
> plot(x3,y)







There is no strong correlation between y and x’s

MODELING NONLINEAR RELATIONS

Develop the model

```
> mymodel = lm(y ~ x1 + x2 + x3, data = mydata)
> summary(mymodel)
```

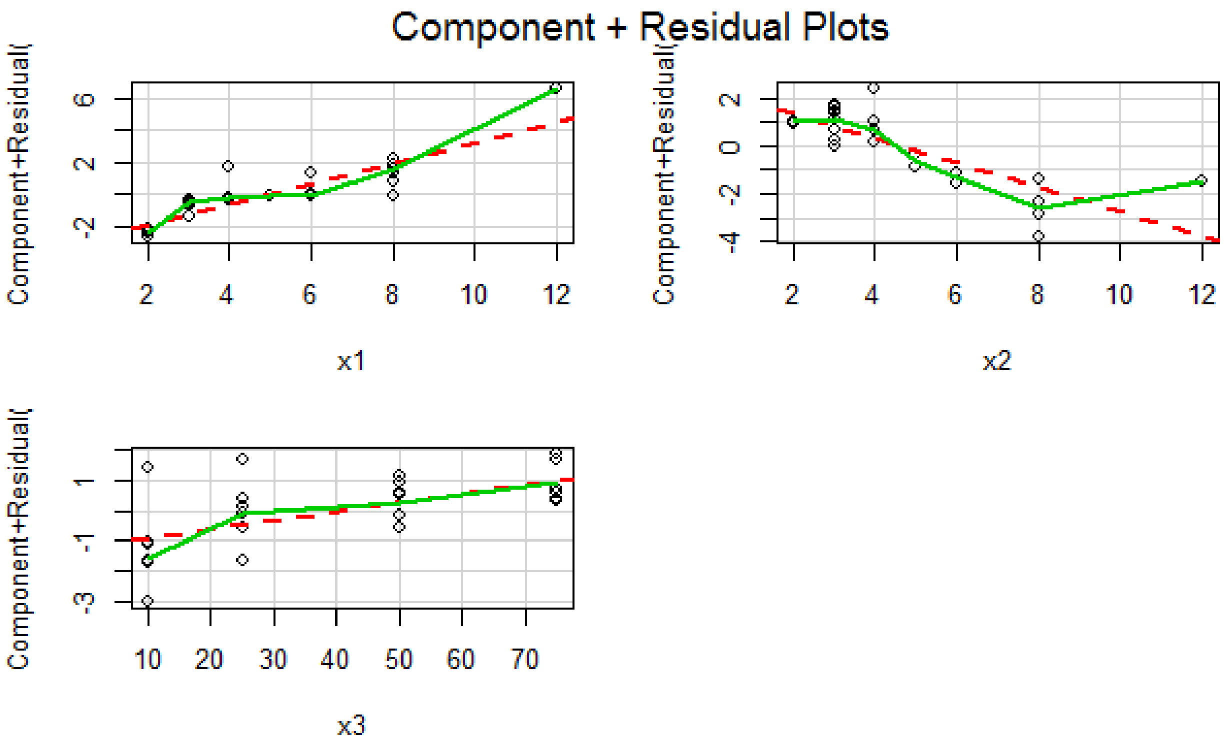
	Estimate	Std. Error	t	p value
(Intercept)	3.58315	0.726839	4.93	0.0001
x1	0.651547	0.0855	7.62	0.0000
x2	-0.509866	0.097132	-5.249	0.0000
x3	0.028888	0.009021	3.202	0.00428

R^2	0.786
Adjusted R^2	0.7563

MODELING NONLINEAR RELATIONS

Develop the model

```
> library(car)
> crPlots(mymodel))
```

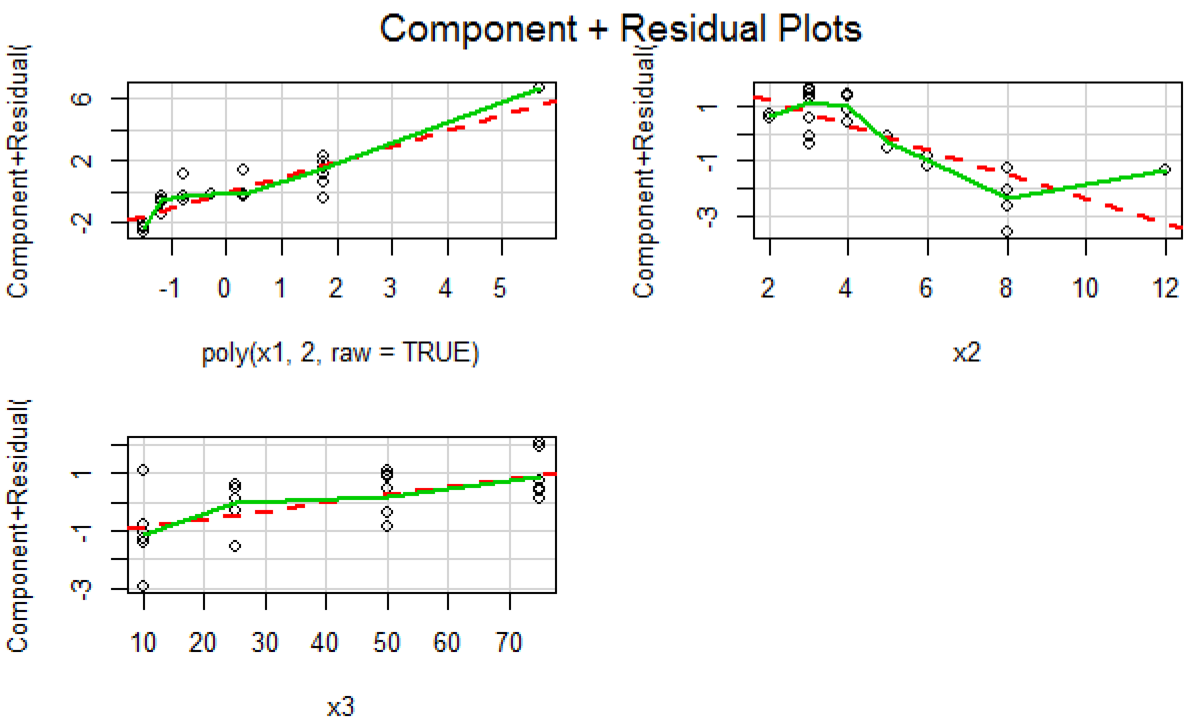


Since the best fit line different from residual line, it is possible improve the model by adding higher order terms

MODELING NONLINEAR RELATIONS

Develop the model

```
> mymodel = lm(y ~ poly(x1, 2, raw = TRUE) + x2 + x3, data = mydata)
> crPlots(mymodel)
```

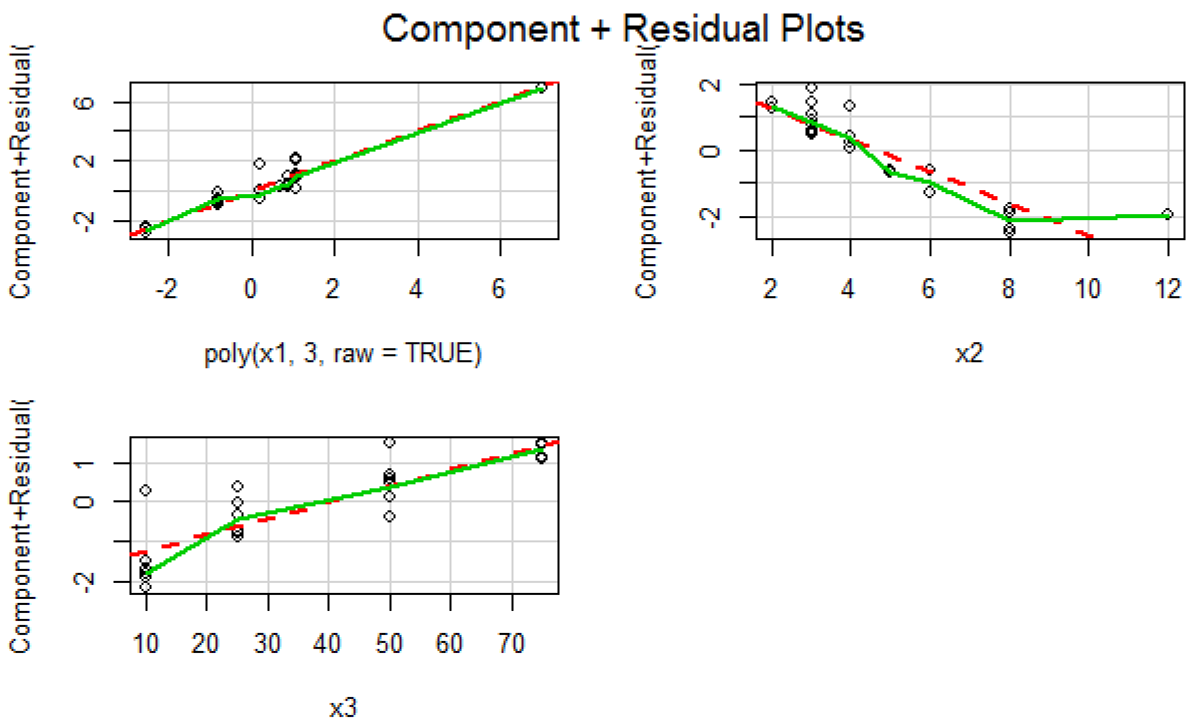


Since the best fit line different from residual line, it is possible improve the model by adding higher order terms

MODELING NONLINEAR RELATIONS

Develop the model

```
> mymodel = lm(y ~ poly(x1, 3, raw = TRUE) + x2 + x3, data = mydata))
> crPlots(mymodel)
```

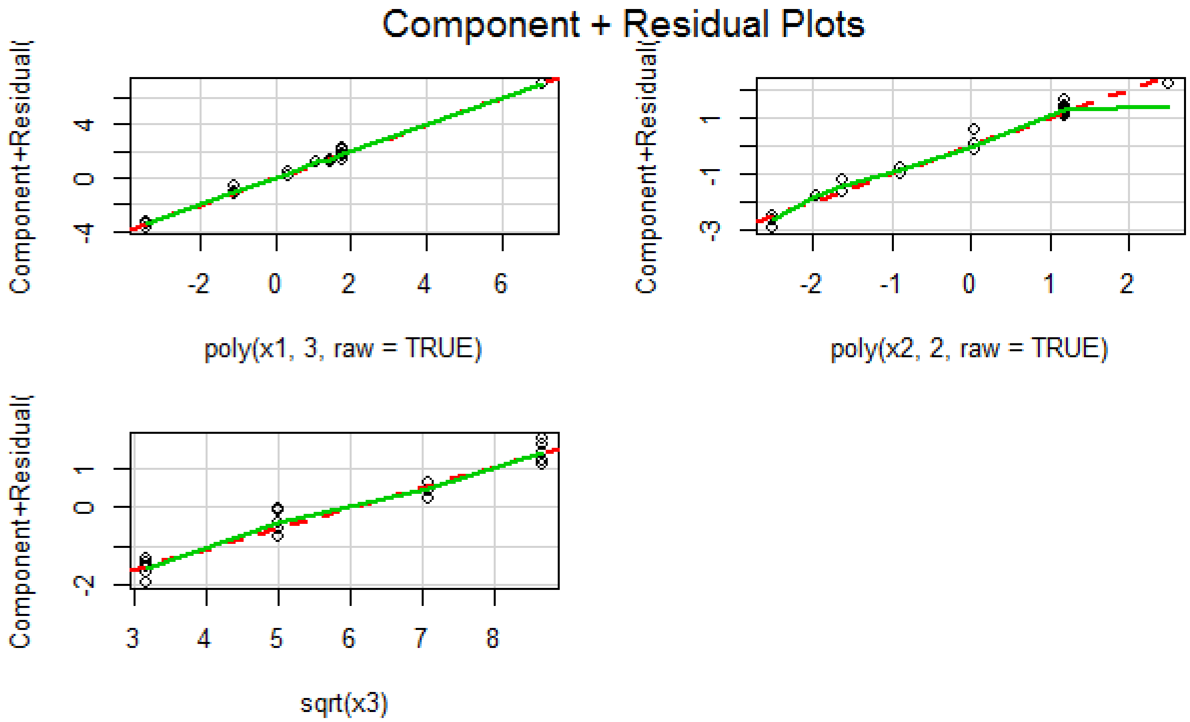


Since the best fit line is more or less overlapping residual line, hence adding square and cube terms of x1 will improve the model. Similarly add additional terms or functions of x2 and x3 to improve the model

MODELING NONLINEAR RELATIONS

Develop the model: Final Model

```
> mymodel = lm(y ~ poly(x1, 3, raw = TRUE) + poly(x2, 2, raw = TRUE) + sqrt(x3), data = mydata))
> crPlots(mymodel)
```



MODELING NONLINEAR RELATIONS

Develop the model: Final Model

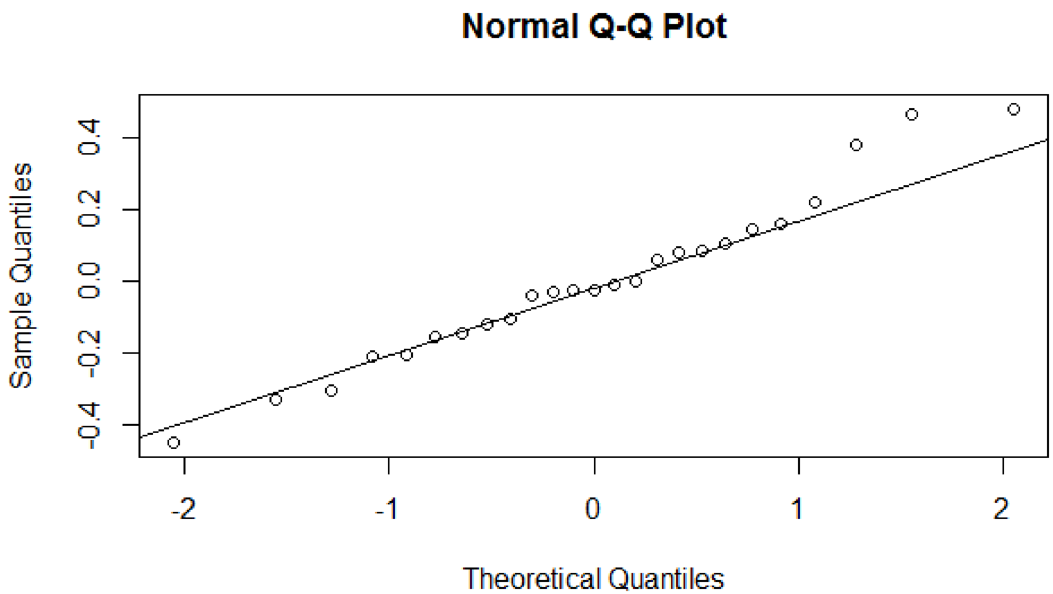
	Estimate	Std. Error	t	p value
(Intercept)	-3.48301	0.705793	-4.935	0.000107
x_1	5.503467	0.36278	15.17	0.0000
x_1^2	-0.77878	0.056814	-13.708	0.0000
x_1^3	0.037516	0.002685	13.971	0.0000
x_2	-1.81437	0.146304	-12.401	0.0000
x_2^2	0.097886	0.010374	9.435	0.0000
$\sqrt{x_3}$	0.527417	0.030664	17.2	0.0000

R^2	0.9881
Adjusted R^2	0.9841

MODELING NONLINEAR RELATIONS

Develop the model: Final Model

```
> res = residuals(mymodel)
> qqnorm(res)
 > qqline(res)
> shapiro.test(res)
```



Shapiro test for Normality	
w	0.9704
p value	0.6569

MODELING NONLINEAR RELATIONS

Exercise 1: Sidewall panel for the interior of an airplane are formed in a 1500 – ton press. The unit manufacturing cost varies with the production lot size. The data shown below give the average cost per unit (in hundreds of dollars) for this product(y) and the production lot size (x). Develop a suitable model for cost in terms of production lot size? The data is given in file Nonlinear_Cost.csv?

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: Develop a model to predict the number of visits of family to a vacation resort based on the salient characteristics of the families. The data collected from 30 households is given in Resort_Visit.csv

1. Reading the file and variables

```
> mydata = Resort_Visit
> visit = mydata$Resort_Visit
> income = mydata$Family_Income
> attitude = mydata$Attitude.Towards.Travel
> importance = mydata$Importance_Vacation
> size = mydata$House_Size
> age = mydata$Age._Head
```

2. Converting response variable to discrete

```
> visit = factor(visit)
```


273

BINARY LOGISTIC REGRESSION

3. Correlation Matrix

```
> cor(mydata)
```

	Resort_Visit	Family_Income	Attitude_Travel	Importance_Vacation	House_Size	Age_Head
Resort_Visit	1.00	-0.60	-0.27	-0.42	-0.59	-0.21
Family_Income	-0.60	1.00	0.30	0.23	0.47	0.21
Attitude_Travel	-0.27	0.30	1.00	0.19	0.15	-0.13
Importance_Vacation	-0.42	0.23	0.19	1.00	0.30	0.11
House_Size	-0.59	0.47	0.15	0.30	1.00	0.09
Age_Head	-0.21	0.21	-0.13	0.11	0.09	1.00

Interpretation: Correlation between X variables should be low

274

BINARY LOGISTIC REGRESSION

4. Converting response variable to discrete

```
> visit = factor(visit)
```

5. Checking relation between Xs and Y

```
> aggregate(income ~visit, FUN = mean)
> aggregate(attitude ~visit, FUN = mean)
> aggregate(importance ~visit, FUN = mean)
> aggregate(size ~visit, FUN = mean)
> aggregate(age ~visit, FUN = mean)
```

Resort_Visit	Mean				
	Family_Income	Attitude_Travel	Importance_Vacation	House_Size	Age_Head
0	58.5200	5.4000	5.8000	4.3333	53.7333
1	41.9133	4.3333	4.0667	2.8000	50.1333

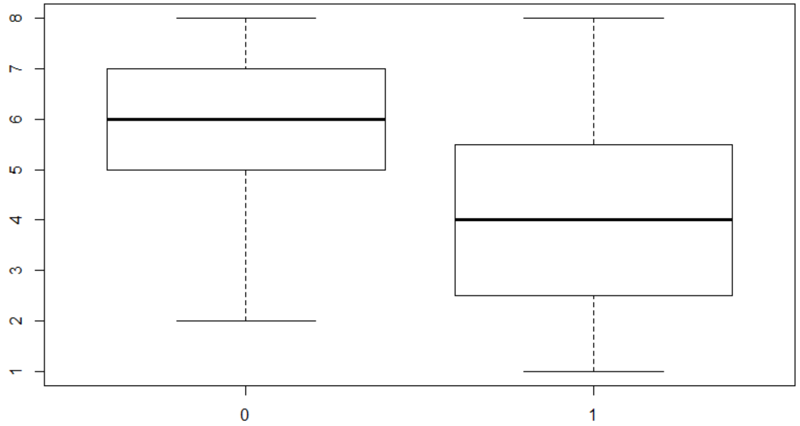
Higher the difference in means, stronger will be the relation to response variable

BINARY LOGISTIC REGRESSION

5. Checking relation between Xs and Y – box plot

```
> boxplot(income ~ visit)
> boxplot(attitude ~ visit)
> boxplot(importance ~ visit)
> boxplot(size ~ visit)
> boxplot(age ~ visit)
```

Income Vs visit



BINARY LOGISTIC REGRESSION

6. Perform Logistic regression

```
> model = glm(visit ~ income + attitude + importance + size + age, family = binomial(logit))

> summary(model)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	15.49503	6.68017	2.32	0.0204
Income	-0.11698	0.06605	-1.771	0.0766
attitude	-0.28129	0.33919	-0.829	0.4069
importance	-0.46157	0.32006	-1.442	0.1493
size	-0.80699	0.49314	-1.636	0.1018
age	-0.07019	0.07199	-0.975	0.3295

BINARY LOGISTIC REGRESSION

6. Perform Logistic regression - ANOVA

```
> anova(model,test = 'Chisq')> summary(model)
```

	Df	Deviance	Resid.Df	Resid.Dev	Pr(>Chi)
NULL	29	41.589			
income	1	12.9813	28	28.608	0.00031
attitude	1	0.4219	27	28.186	0.51598
importance	1	3.8344	26	24.351	0.05021
size	1	3.4398	25	20.911	0.06364
age	1	1.0242	24	19.887	0.31152

Since p value < 0.05 for Income, Importance_Vacation & Size, redo the modelling with important factors only

278

BINARY LOGISTIC REGRESSION

7. Perform Logistic regression - Modified

	Estimate	Std Error	z value	p value
(Intercept)	8.46599	3.02494	2.799	0.00513
Income	-0.10641	0.05156	-2.064	0.03904
Size	-0.93539	0.47632	-1.964	0.04955

Since p value < 0.05 for both factors, Income & Size, the response variable can be modelled in terms of those two factors

The model is

$$y = \frac{e^{8.46599 - 0.10641 \text{Annual\_Income} - 0.93539 Size}}{1 + e^{8.46599 - 0.10641 \text{Annual\_Income} - 0.93539 Size}}$$

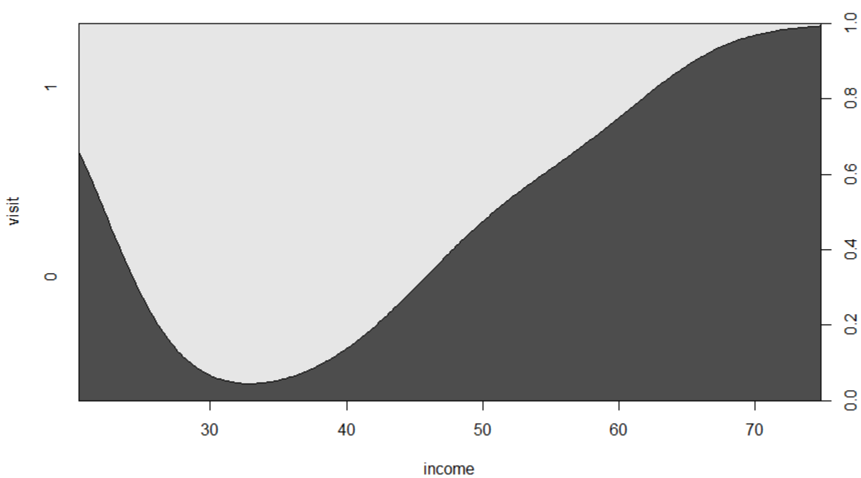
279

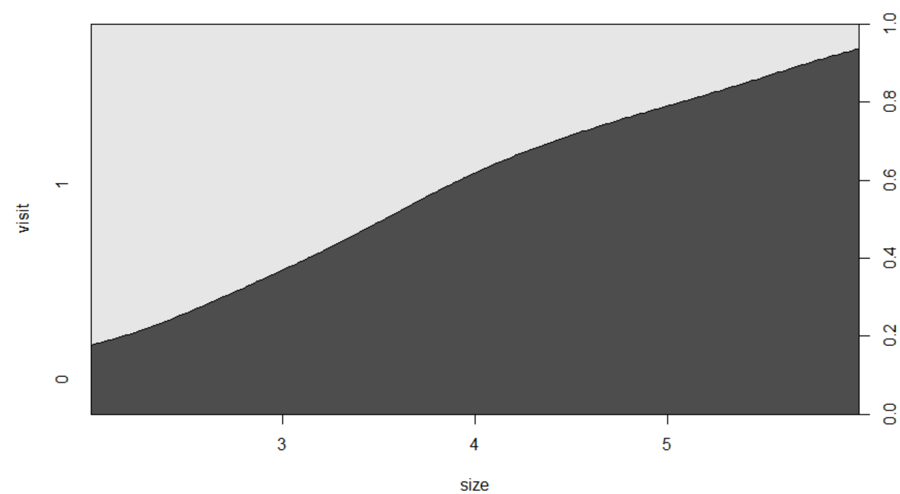
8. Conditional Density plots (Response Vs Factors)

Describing how the conditional distribution of a categorical variable y changes over a numerical variable x

> cdplot(visit ~ income)

> cdplot(visit ~ size)





280

BINARY LOGISTIC REGRESSION

9. Fitted Values and residuals

```
> predict(model,type = 'response')
> residuals(model,type = 'deviance')
> predclass = ifelse(predict(model, type ='response')>0.5,"1","0")
```

SL No.	Actual	Fitted	Residuals	Predicted Class	SL No.	Actual	Fitted	Residuals	Predicted Class
1	0	0.970979	-2.66073	1	16	1	0.904132	0.448954	1
2	0	0.059732	-0.35097	0	17	1	0.939523	0.353222	1
3	0	0.021049	-0.20627	0	18	1	0.880611	0.50426	1
4	0	0.202309	-0.67236	0	19	1	0.345537	1.457845	0
5	0	0.292461	-0.83182	0	20	1	0.724535	0.802777	1
6	0	0.014893	-0.17324	0	21	1	0.925508	0.393479	1
7	0	0.677783	-1.50501	1	22	1	0.677559	0.882337	1
8	0	0.038723	-0.28105	0	23	1	0.680103	0.878079	1
9	0	0.109432	-0.48145	0	24	1	0.516151	1.150092	1
10	0	0.030543	-0.24908	0	25	1	0.680326	0.877704	1
11	0	0.017609	-0.1885	0	26	1	0.77062	0.721887	1
12	0	0.050856	-0.32309	0	27	1	0.629425	0.962235	1
13	0	0.04202	-0.29301	0	28	1	0.954395	0.305541	1
14	0	0.601981	-1.35739	1	29	1	0.841493	0.587498	1
15	0	0.499424	-1.17643	0	30	1	0.900286	0.45835	1

281

10. Model Evaluation

```
> mytable = table(visit, predclass)
> mytable
> prop.table(mytable)
```

	Predicted Count		Total
Actual Count	0	1	
0	12	3	15
1	1	14	15
Total	13	17	30

	Predicted %		Total
Actual %	0	1	
0	40	10	50
1	3	47	50
Total	43	50	100

Statistics	Value
Accuracy %	87
Error %	13

Accuracy of ≥ 80 % is good

282

BINARY LOGISTIC REGRESSION

Exercise 2: A car rental company wants to develop a model for brand loyalty. The data was collected from 30 customers, 15 of whom are brand loyal (indicated by 1) and 15 of whom are not (indicated by 0). The company also measured attitude towards the brand (Brand), attitude towards the type of vehicle (vehicle) and attitude toward availing rent a car service (Service), all on a 1 (unfavorable) to 7 (favorable) scale. The data is given in brand.csv file.

TREE BASED METHODS

CLASSIFICATION AND REGRESSION TREE

Objective

To develop a predictive model to classify dependant or response metric (Y) in terms of independent or exploratory variables(Xs).

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 discrete

Method = “anova”

CLASSIFICATION AND REGRESSION TREE

Classifies data (develops a model) based on the training data

Each sample is assumed to belong to a predefined class

Sample data set used for building the model is training set

Usage:

For classifying future or unknown data

CLASSIFICATION AND REGRESSION TREE

Example:

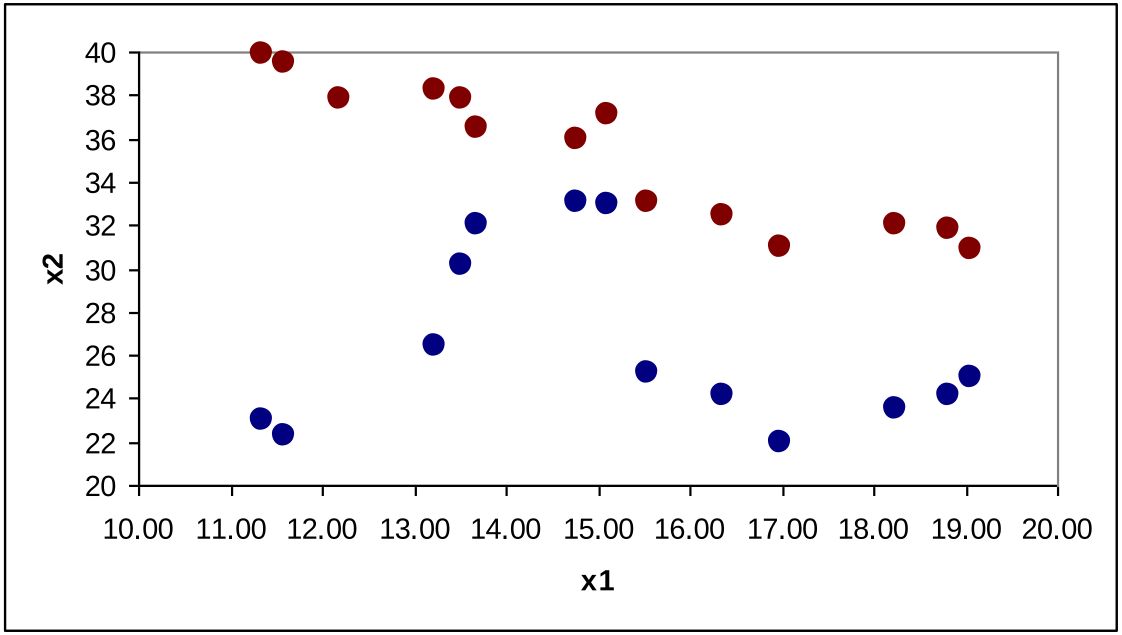
Attribute 1	x1
Attribute 2	x2
Label : y	Y1 (Red) , y2 (Blue)

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:

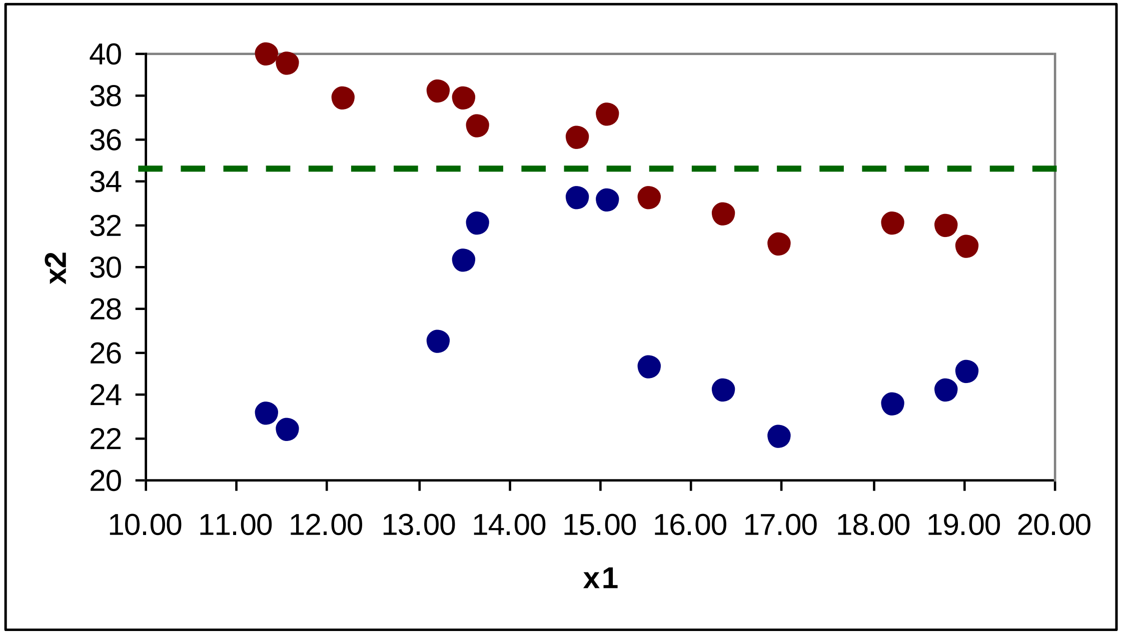
Attribute 1	x1
Attribute 2	x2
Label : y	Y1 (Red) , y2 (Blue)

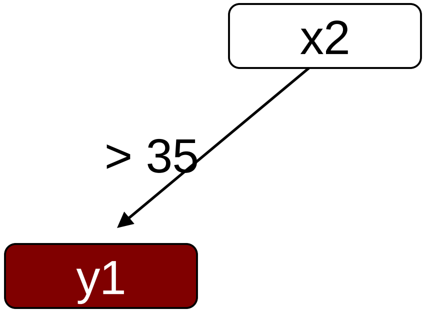


CLASSIFICATION AND REGRESSION TREE

Example:

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

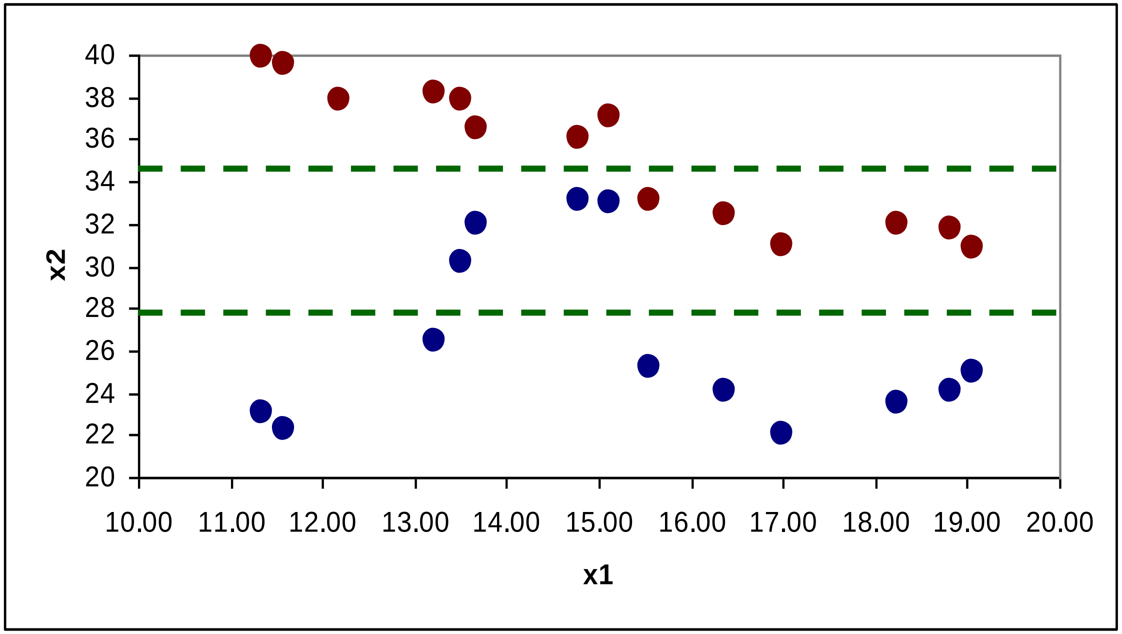


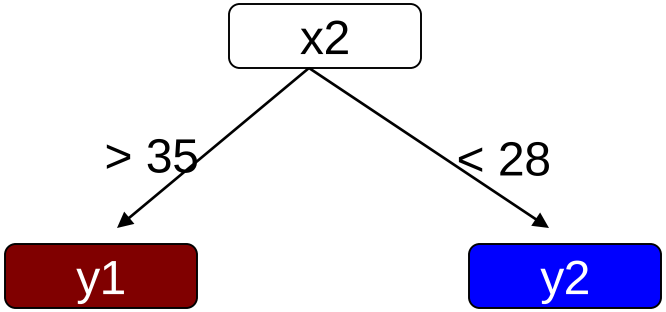


CLASSIFICATION AND REGRESSION TREE

Example:

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

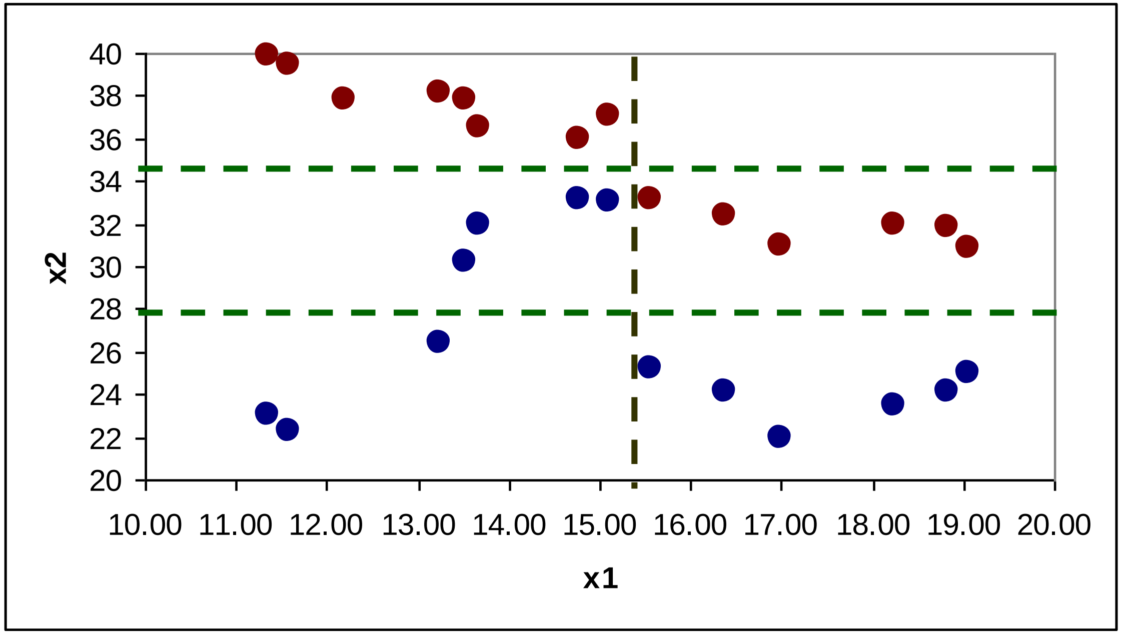


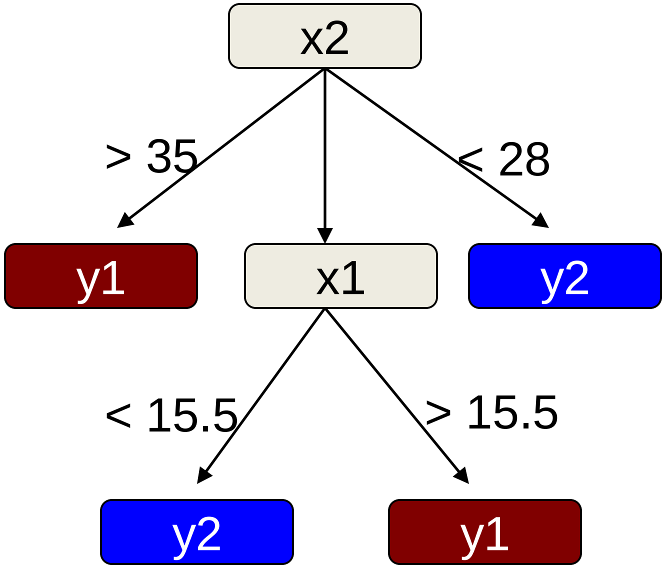


CLASSIFICATION AND REGRESSION TREE

Example:

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

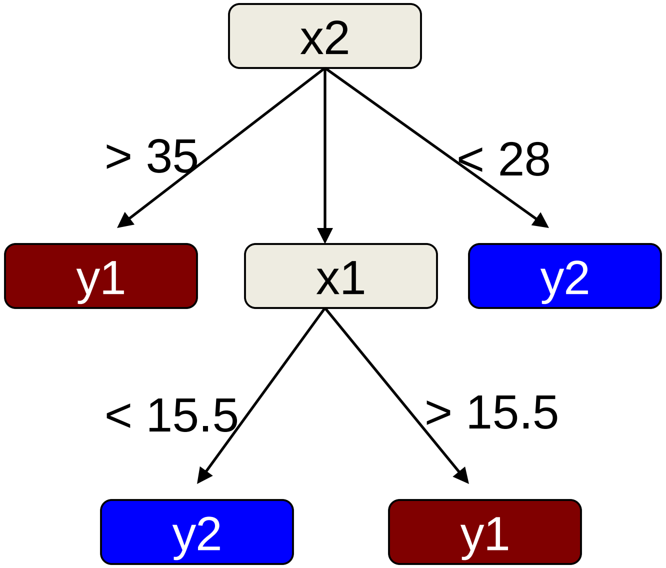




CLASSIFICATION AND REGRESSION TREE

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

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

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?

SL No	District	House Type	Income	Previous_Customer	Outcome
1	Suburban	Detached	High	No	No Response
2	Suburban	Detached	High	Yes	No Response
3	Rural	Detached	High	No	Responded
4	Urban	Semi-detached	High	No	Responded
5	Urban	Semi-detached	Low	No	Responded
6	Urban	Semi-detached	Low	Yes	No Response
7	Rural	Semi-detached	Low	Yes	Responded
8	Suburban	Terrace	High	No	No Response
9	Suburban	Semi-detached	Low	No	Responded
10	Urban	Terrace	Low	No	Responded
11	Suburban	Terrace	Low	Yes	Responded
12	Rural	Terrace	High	Yes	Responded
13	Rural	Detached	Low	No	Responded
14	Urban	Terrace	High	Yes	No Response

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?

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)
> 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”

```
>print(mymodel)
```

CLASSIFICATION AND REGRESSION TREE

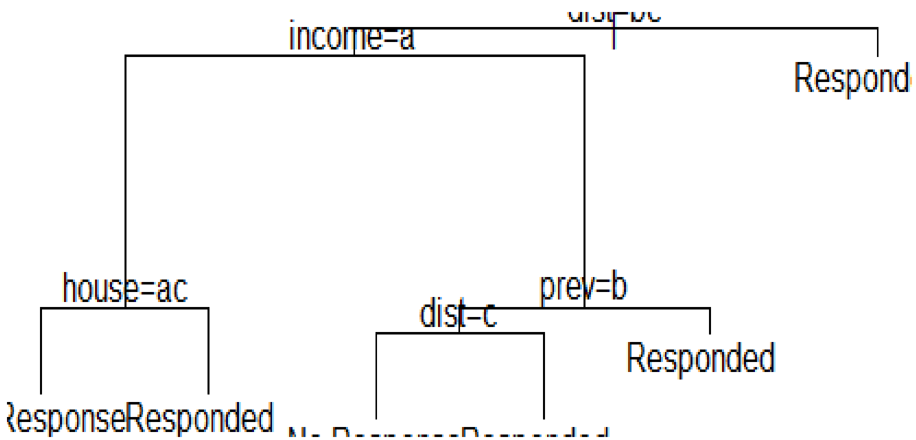
```
1) root 14 5 Responded (0.3571429 0.6428571)
  2) dist=Suburban,Urban 10 5 No Response (0.5000000 0.5000000)
    4) income=High 5 1 No Response (0.8000000 0.2000000)
      8) house=Detached,Terrace 4 0 No Response (1.0000000 0.0000000) *
      9) house=Semi-detached 1 0 Responded (0.0000000 1.0000000) *

    5) income=Low 5 1 Responded (0.2000000 0.8000000)
      10) prev=Yes 2 1 No Response (0.5000000 0.5000000)
        20) dist=Urban 1 0 No Response (1.0000000 0.0000000) *
        21) dist=Suburban 1 0 Responded (0.0000000 1.0000000) *

      11) prev=No 3 0 Responded (0.0000000 1.0000000) *

  3) dist=Rural 4 0 Responded (0.0000000 1.0000000) *
```

CLASSIFICATION AND REGRESSION TREE

Plot the tree

> plot(mymodel)

> text(mymodel)



CLASSIFICATION AND REGRESSION TREE

Making predictions

> pred = predict(mymodel)

> Predclass = ifelse(pred[,1] > 0.5, “1”, “2”)

> mytable = table(outcome, predclass)

		Predicted	
		Respond	No Respond
Outcome	Respond	9	0
	No Respond	0	5

CLASSIFICATION AND REGRESSION TREE

Exercise 1: Develop a tree based model for predicting whether the customer will take pep using ghe customer profile data given in bank-data.csv?

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?

MULTI RESPONSE SCORING METHODS

Based on

1. Taguchi’s Loss Function Approach
2. Derringer’s Desirability Function Approach

Taguchi's Loss Function Approach

Types of Metrics / Variables

a. Larger the better

Eg: % Utilization, CPE, Productivity, Mileage

Target : 100% or Infinity

b. Smaller the better

Eg: IRT, TMPI, DTS, etc.

Target: 0

c. Nominal the better

Eg: Number of Cases Created, Weight, Dimensions, etc.

Target: A specified value T

Taguchi's Loss Function Approach

Example: The data on the performance of 10 clusters based on IRT, Utilization, CPE and cases created are given below. The values of target, upper specification limit (USL), lower specification limit (LSL) is also given. Rate the clusters using Taguchi's Loss Function.

			CPE		Cases
Cluster	IRT	Utilization	Bottom Box	Top Box	Created
1	1.5	92	4.5	70.5	279
2	0.7	85	2.3	85.7	259
3	1.2	93	6.2	68.8	128
4	2	71	0.2	95.8	129
5	2.5	84	1.8	92.2	279
6	0.8	85	4.2	87.8	202
7	1.4	65	6.3	78.7	260
8	1.5	93	3.4	80.6	142
9	1.2	96	3.6	81.4	166
10	1.3	79	5.2	81.8	235
Target	0	100	0	100	200
USL	2		5		300
LSL		55		75	100

Taguchi's Loss Function Approach

Taguchi's Loss Function

$$L(\text{Value}) = k(\text{value} - T)^2$$

Where

T: Target

k: Quality loss coefficient

Note:

1. Loss L(value) = 0 when value is on target
2. Choose k such that loss L(value) = 1, when value is on specification limits

Taguchi's Loss Function Approach

Taguchi's Loss Function

$$L(value) = k(value - T)^2$$

1. Smaller the better type

Target = 0, k = 1 / USL^2

$$L(value) = \frac{value^2}{USL^2}$$

2. Larger the better type

Target = ∞, k = 1 / LSL^2

$$L(value) = \frac{1}{(1/LSL)^2}\frac{1}{value^2}$$

3. Nominal the best type

Target = t, k = 4 / (USL – LSL)^2

$$L(y) = \frac{4}{(USL - LSL)^2}(value - T)^2$$

Taguchi's Loss Function Approach

Step 1: Convert larger the better type variables into smaller the better type

Cluster	IRT	1/Utilization	CPE		Cases Created
			Bottom Box	1/Top Box	
1	1.5	0.0109	4.5	0.0142	279
2	0.7	0.0118	2.3	0.0117	259
3	1.2	0.0108	6.2	0.0145	128
4	2	0.0141	0.2	0.0104	129
5	2.5	0.0119	1.8	0.0108	279
6	0.8	0.0118	4.2	0.0114	202
7	1.4	0.0154	6.3	0.0127	260
8	1.5	0.0108	3.4	0.0124	142
9	1.2	0.0104	3.6	0.0123	166
10	1.3	0.0127	5.2	0.0122	235

Target	0	0	0	0	200
USL	2	0.01818	5	0.0133	300
LSL					100

Taguchi's Loss Function Approach

Step 2: Calculate the Loss function for each variable

Cluster	IRT	Utilization	CPE		Cases Created
			Bottom Box	Top Box	
1	0.5625	0.3574	0.8100	1.1317	0.6241
2	0.1225	0.4187	0.2116	0.7659	0.3481
3	0.3600	0.3498	1.5376	1.1884	0.5184
4	1.0000	0.6001	0.0016	0.6129	0.5041
5	1.5625	0.4287	0.1296	0.6617	0.6241
6	0.1600	0.4187	0.7056	0.7297	0.0004
7	0.4900	0.7160	1.5876	0.9082	0.3600
8	0.5625	0.3498	0.4624	0.8659	0.3364
9	0.3600	0.3282	0.5184	0.8489	0.1156
10	0.4225	0.4847	1.0816	0.8407	0.1225

Taguchi's Loss Function Approach

Step 3: Calculate the Overall expected loss

Overall Expected Loss = Average of individual Loss functions

Cluster	IRT	Utilization	CPE		Cases Created	Expected Loss
			Bottom Box	Top Box		
1	0.5625	0.3574	0.8100	1.1317	0.6241	0.6971
2	0.1225	0.4187	0.2116	0.7659	0.3481	0.3734
3	0.3600	0.3498	1.5376	1.1884	0.5184	0.7908
4	1.0000	0.6001	0.0016	0.6129	0.5041	0.5437
5	1.5625	0.4287	0.1296	0.6617	0.6241	0.6813
6	0.1600	0.4187	0.7056	0.7297	0.0004	0.4029
7	0.4900	0.7160	1.5876	0.9082	0.3600	0.8124
8	0.5625	0.3498	0.4624	0.8659	0.3364	0.5154
9	0.3600	0.3282	0.5184	0.8489	0.1156	0.4342
10	0.4225	0.4847	1.0816	0.8407	0.1225	0.5904

Taguchi's Loss Function Approach

Step 4: Rank the items in the descending order of overall loss value

Cluster	IRT	Utilization	CPE		Cases Created	Expected Loss	Rank
			Bottom Box	Top Box			
1	0.5625	0.3574	0.8100	1.1317	0.6241	0.6971	8
2	0.1225	0.4187	0.2116	0.7659	0.3481	0.3734	1
3	0.3600	0.3498	1.5376	1.1884	0.5184	0.7908	9
4	1.0000	0.6001	0.0016	0.6129	0.5041	0.5437	5
5	1.5625	0.4287	0.1296	0.6617	0.6241	0.6813	7
6	0.1600	0.4187	0.7056	0.7297	0.0004	0.4029	2
7	0.4900	0.7160	1.5876	0.9082	0.3600	0.8124	10
8	0.5625	0.3498	0.4624	0.8659	0.3364	0.5154	4
9	0.3600	0.3282	0.5184	0.8489	0.1156	0.4342	3
10	0.4225	0.4847	1.0816	0.8407	0.1225	0.5904	6

Taguchi's Loss Function Approach

Exercise: Rate the clusters based on the following parameters

Cluster	Vertical	Region	IRT	TMPI	Utilization	DTC
1	EOS	US	7.6	2.5	73.9	9.8
2	EOS	EMEA	1.9	3	71.5	2.6
3	EOS	India	7.1	2.1	26.2	5.4
4	ECS	US	0.5	0.2	49.1	3.7
5	ECS	EMEA	0.5	3.2	92.3	3.5
6	ECS	India	8.1	0.7	88.9	6
7	DS	US	3.3	0.7	84.9	5.1
8	DS	EMEA	5.1	1.5	36.7	7.5
9	DS	India	4.8	3.2	61	4.5
10	EPS	US	2.9	0.6	75.5	1.5
11	EPS	EMEA	3.4	3.2	72.3	2.1
12	EPS	India	5.5	1.5	84	3.4

Target	0	0	100	0
USL	8	2		10
LSL			75	

Derringer's Desirability Function Approach

Example: The data on the performance of 10 clusters based on IRT, Utilization, CPE and cases created are given below. The values of target, upper specification limit (USL), lower specification limit (LSL) is also given. Rate the clusters using Desirability Function.

Cluster	IRT	Utilization	CPE		Cases Created
			Bottom Box	Top Box	
1	1.5	92	4.5	70.5	279
2	0.7	85	2.3	85.7	259
3	1.2	93	6.2	68.8	128
4	2	71	0.2	95.8	129
5	2.5	84	1.8	92.2	279
6	0.8	85	4.2	87.8	202
7	1.4	65	6.3	78.7	260
8	1.5	93	3.4	80.6	142
9	1.2	96	3.6	81.4	166
10	1.3	79	5.2	81.8	235

Target	0	100	0	100	200
USL	2		5		300
LSL		55		75	100

Derringer's Desirability Function Approach

Desirability Function

1. Nominal the Best

If value is between LSL and Target

$$d = \left| \frac{Value - LSL}{Target - LSL} \right|^{0.5}$$

Else if value is between USL and Target

$$d = \left| \frac{Value - USL}{Target - USL} \right|^{0.5}$$

d = 0, otherwise

Derringer's Desirability Function Approach

Desirability Function

2. Smaller the Better

If value is between USL and $\text{Value}_{\text{minimum}}$

$$d = \left| \frac{Value - USL}{Value_{\text{minimum}} - USL} \right|^{0.5}$$

d = 0, if value > USL

d = 1. If value < $\text{Value}_{\text{minimum}}$

$\text{Value}_{\text{minimum}}$ is the minimum possible value

Derringer’s Desirability Function Approach

Desirability Function

3. Larger the Better

If value is between USL and $\text{Value}_{\text{maximum}}$

$$d = \left| \frac{Value - LSL}{Value_{maximum} - LSL} \right|^{0.5}$$

d = 0, if value L LSL

D = 1. If value > $\text{Value}_{\text{maximum}}$

$\text{Value}_{\text{maximum}}$ is the maximum possible value

Derringer's Desirability Function Approach

Desirability Function

Overall Desirability

D = Geometric mean of individual desirability values

If there are p variables with desirability values d_1, d_2, - - - , d_p , then

Overall Desirability

$$D = (d_1 \times d_2 \times --- \times d_p)^{1/p}$$

Note: d = 1, if value is on target

Derringer's Desirability Function Approach

Step 1: Identify the Minimum for smaller the better and Maximum for larger the better

Cluster	IRT	Utilization	CPE		Cases Created
			Bottom Box	Top Box	
1	1.5	92	4.5	70.5	279
2	0.7	85	2.3	85.7	259
3	1.2	93	6.2	68.8	128
4	2	71	0.2	95.8	129
5	2.5	84	1.8	92.2	279
6	0.8	85	4.2	87.8	202
7	1.4	65	6.3	78.7	260
8	1.5	93	3.4	80.6	142
9	1.2	96	3.6	81.4	166
10	1.3	79	5.2	81.8	235

Target	0	100	0	100	200
USL	2		5		300
LSL		55		75	100
Minimum	0		0		
Maximum		100		100	

Derringer's Desirability Function Approach

Step 2: Calculate the desirability function for each variable

Cluster	IRT	Utilization	CPE		Cases Created
			Bottom Box	Top Box	
1	0.6202	0.9500	0.3227	0.0000	0.4583
2	1.0000	0.8554	0.7500	0.7172	0.6403
3	0.7845	0.9627	0.0000	0.0000	0.5292
4	0.0000	0.6247	1.0000	1.0000	0.5385
5	0.0000	0.8410	0.8165	0.9094	0.4583
6	0.9608	0.8554	0.4082	0.7845	0.9899
7	0.6794	0.4939	0.0000	0.4218	0.6325
8	0.6202	0.9627	0.5774	0.5189	0.6481
9	0.7845	1.0000	0.5401	0.5547	0.8124
10	0.7338	0.7651	0.0000	0.5718	0.8062

Derringer's Desirability Function Approach

Step 3: Calculate the Overall Desirability Function

Overall Desirability = Geometric mean of individual desirability functions

Cluster	IRT	Utilization	CPE		Cases Created	Overall Desirability
			Bottom Box	Top Box		
1	0.6202	0.9500	0.3227	0.0000	0.4583	0.0000
2	1.0000	0.8554	0.7500	0.7172	0.6403	0.7832
3	0.7845	0.9627	0.0000	0.0000	0.5292	0.0000
4	0.0000	0.6247	1.0000	1.0000	0.5385	0.0000
5	0.0000	0.8410	0.8165	0.9094	0.4583	0.0000
6	0.9608	0.8554	0.4082	0.7845	0.9899	0.7642
7	0.6794	0.4939	0.0000	0.4218	0.6325	0.0000
8	0.6202	0.9627	0.5774	0.5189	0.6481	0.6499
9	0.7845	1.0000	0.5401	0.5547	0.8124	0.7181
10	0.7338	0.7651	0.0000	0.5718	0.8062	0.0000

Derringer's Desirability Function Approach

Step 4: Rank the items in the descending order of overall loss value

Cluster	IRT	Utilization	CPE		Cases Created	Overall Desirability	Rank
			Bottom Box	Top Box			
1	0.6202	0.9500	0.3227	0.0000	0.4583	0.0000	5
2	1.0000	0.8554	0.7500	0.7172	0.6403	0.7832	1
3	0.7845	0.9627	0.0000	0.0000	0.5292	0.0000	5
4	0.0000	0.6247	1.0000	1.0000	0.5385	0.0000	5
5	0.0000	0.8410	0.8165	0.9094	0.4583	0.0000	5
6	0.9608	0.8554	0.4082	0.7845	0.9899	0.7642	2
7	0.6794	0.4939	0.0000	0.4218	0.6325	0.0000	5
8	0.6202	0.9627	0.5774	0.5189	0.6481	0.6499	4
9	0.7845	1.0000	0.5401	0.5547	0.8124	0.7181	3
10	0.7338	0.7651	0.0000	0.5718	0.8062	0.0000	5

Derringer's Desirability Function Approach

Exercise: Rate the clusters based on the following parameters

Cluster	Vertical	Region	IRT	TMPI	Utilization	DTC
1	EOS	US	7.6	2.5	73.9	9.8
2	EOS	EMEA	1.9	3	71.5	2.6
3	EOS	India	7.1	2.1	26.2	5.4
4	ECS	US	0.5	0.2	49.1	3.7
5	ECS	EMEA	0.5	3.2	92.3	3.5
6	ECS	India	8.1	0.7	88.9	6
7	DS	US	3.3	0.7	84.9	5.1
8	DS	EMEA	5.1	1.5	36.7	7.5
9	DS	India	4.8	3.2	61	4.5
10	EPS	US	2.9	0.6	75.5	1.5
11	EPS	EMEA	3.4	3.2	72.3	2.1
12	EPS	India	5.5	1.5	84	3.4

Target	0	0	100	0
USL	8	2		10
LSL			75	

MARKET BASKET ANALYSIS

MARKET BASKET ANALYSIS

A modeling technique based upon the logic that if a customer buy a certain group of items, he is more (or less) likely to buy another group of items

Example:

Those who buy cigarettes are more likely to buy match box also.

MARKET BASKET ANALYSIS

Association Rule Mining:

Developing rules that predict the occurrence of of an item based on the occurrence of other items in the transaction

Example

Id	Items
1	Milk, Bread
2	Bread, Biscuits, Toys, Eggs
3	Milk, Biscuits, Toys, Fruits
4	Bread, Milk, Toys, Biscuits
5	Milk, Bread, Biscuits, Fruits

{Milk, Bread} → {Biscuits} with probability = 2 / 3

MARKET BASKET ANALYSIS

Itemset:

A collection of one or more items

k – itemset

An itemset consisting of k items

Id	Items
1	Milk, Bread
2	Bread, Biscuits, Toys, Eggs
3	Milk, Biscuits, Toys, Fruits
4	Bread, Milk, Toys, Biscuits
5	Milk, Bread, Biscuits, Fruits

MARKET BASKET ANALYSIS

Support count:

Frequency of occurrence of an itemset

Example

{Milk, Bread, Biscuits} = 2

Id	Items
1	Milk, Bread
2	Bread, Biscuits, Toys, Eggs
3	Milk, Biscuits, Toys, Fruits
4	Bread, Milk, Toys, Biscuits
5	Milk, Bread, Biscuits, Fruits

MARKET BASKET ANALYSIS

Support :

Proportion or fraction of transaction that contain an itemset

Example

{Milk, Bread, Biscuits} = 2 / 5

Id	Items
1	Milk, Bread
2	Bread, Biscuits, Toys, Eggs
3	Milk, Biscuits, Toys, Fruits
4	Bread, Milk, Toys, Biscuits
5	Milk, Bread, Biscuits, Fruits

Frequent Itemset

An itemset whose support is greater than or equal to minimum support

MARKET BASKET ANALYSIS

Confidence

Conditional probability that an item will appear in transactions that contain another items

Example

Confidence that Toys will appear in transaction containing Milk & Biscuits

= {Milk, Biscuits, Toys} / {Milk, Biscuits} = 2 / 3 = 0.67

Id	Items
1	Milk, Bread
2	Bread, Biscuits, Toys, Eggs
3	Milk, Biscuits, Toys, Fruits
4	Bread, Milk, Toys, Biscuits
5	Milk, Bread, Biscuits, Fruits

MARKET BASKET ANALYSIS

Association Rule Mining

1. Frequent Itemset Generation

   Fix minimum support value

   Generate all itemsets whose support ≥ minimum support

2. Rule Generation

   Fix minimum confidence value

   Generate high confidence rules from each frequent itemset

Frequent Itemset Generation: Apriori Algorithm

a. Fix minimum support count

b. Generate all itemsets of length = 1

c. Calculate the support for each itemset

d. Eliminate all itemsets with support count < minimum support count

e. Repeat steps c & d for itemsets of length = 2, 3, ---

MARKET BASKET ANALYSIS

Frequent Itemset Generation: Apriori Algorithm

Example:

Minimum Support count = 2

Id	Items
1	A,C,D
2	B,C,E
3	A,B,C,E
4	B,E
5	A,E
6	A,C,E

Frequent Itemset Generation: Apriori Algorithm

Example:

Minimum Support count = 2

Step 1:

Generate itemsets of length = 1 & calculate support

Item	Support count
A	4
B	3
C	4
D	1
E	5

MARKET BASKET ANALYSIS

Frequent Itemset Generation: Apriori Algorithm

Example:

Minimum Support count = 2

Step 2:

eliminate itemsets with support count < minimum support count (2)

Item	Support count
A	4
B	3
C	4
D	1
E	5

MARKET BASKET ANALYSIS

Frequent Itemset Generation: Apriori Algorithm

Example:

Minimum Support count = 2

Step 2:

eliminate itemsets with support count < minimum support count (2)

Item	Support count
A	4
B	3
C	4
E	5

MARKET BASKET ANALYSIS

Frequent Itemset Generation: Apriori Algorithm

Example:

Minimum Support count = 2

Step 3:

generate itemsets of length = 2

Item	Support count
A, B	1
A, C	3
A,E	3
B, C	2
B, E	3
C,E	3

MARKET BASKET ANALYSIS

Frequent Itemset Generation: Apriori Algorithm

Example:

Minimum Support count = 2

Step 4:

eliminate itemsets with support count < minimum support count (2)

Item	Support count
A, B	1
A, C	3
A,E	3
B, C	2
B, E	3
C,E	3

Frequent Itemset Generation: Apriori Algorithm

Example:

Minimum Support count = 2

Step 4:

eliminate itemsets with support count < minimum support count (2)

Item	Support count
A, C	3
A,E	3
B, C	2
B, E	3
C,E	3

MARKET BASKET ANALYSIS

Frequent Itemset Generation: Apriori Algorithm

Example:

Minimum Support count = 2

Step 5:

generate itemsets of length = 3

Item	Support count
A, C, E	2
B, C, E	2

MARKET BASKET ANALYSIS

Frequent Itemset Generation: Apriori Algorithm

Example:

Minimum Support count = 2

Step 6:

generate itemsets of length = 4

Itemset	Support Count
A, B, C, E	1

MARKET BASKET ANALYSIS

Frequent Itemset Generation: Apriori Algorithm

Example:

Minimum Support count = 2

Result:

Item	Support count	Support
A, C, E	2	0.33
B, C, E	2	0.33
A , C	3	0.50
A , E	3	0.50
B,C	2	0.33
B,E	3	0.50
C,E	3	0.50

MARKET BASKET ANALYSIS

Association Rule Mining: Apriori Algorithm

Example:

Minimum Support = 0.50

Minimum Confidence = 0.5

Item	Support count	Support
A, C, E	2	0.33
B, C, E	2	0.33
A , C	3	0.50
A , E	3	0.50
B,C	2	0.33
B,E	3	0.50
C,E	3	0.50

MARKET BASKET ANALYSIS

Association Rule Mining: Apriori Algorithm

Example:

Minimum Support = 0.50

Minimum Confidence = 0.5

Item	Support	Confidence
A → C	0.50	0.75
A → E	0.50	0.75
B → E	0.50	1.00
C → E	0.50	0.75
C → A	0.50	0.75
E → A	0.50	0.60
E → B	0.50	0.60
E → C	0.50	0.60

MARKET BASKET ANALYSIS

Association Rule Mining: Other Measures

Lift

Lift (A $\rightarrow$ C) = Confidence (A $\rightarrow$ C) / Support (C)

Example

Item	Confidence	Support	Lift
A $\rightarrow$ C	0.75	C = 0.67	1.12
A $\rightarrow$ E	0.75	E = 0.83	0.93

Criteria : Lift $\geq$ 1

Lift (A , C) = 1.12 > Lift (A , E) indicates that A has a greater impact on the frequency of C than it has on the frequency of E

MARKET BASKET ANALYSIS

R code

Read the data fie to my data and specify the variables

```
>target = mydata$items
>ident = mydata$Id
```

Make transaction varibale

```
>transactions = as(split(target, ident), “transactions”)
```

Generate Rules

```
>myrules = apriori(transactions, parameter = list(support = 0.25, confidence = 0.50, minlen = 2))
```

MARKET BASKET ANALYSIS

R code

Display rules

```
>myrules

>inspect(myrules)
```

MARKET BASKET ANALYSIS

Exercise 1:

The market basket Software data set contains the details of transaction at a software product company.

1. Identify the frequent product types with a support of minimum 25% ?
2. Also identify the association of products with a confidence of minimum 50% ?
3. What is the chance that Operating System and Office Suite will be purchased together?
4. What is the chance that Operating System and Visual Studio will be purchased together?
5. Estimate the chance that the customers who buy Operating System will also purchase Office Suite ?
6. Estimate the chance that the customers who buy Operating System will also purchase Visual Studio?

FACTOR ANALYSIS

FACTOR ANALYSIS

- A dimensionality reduction technique
- Large number of correlated variables can be reduced to a manageable number of uncorrelated or independent factors.
- The emphasis is on the identification of underlying factors that might explain the dimensions associated with large data sets

$$F_i = w_{i1}x_1 + w_{i2}x_2 + w_{i3}x_3 + ---- + w_{ik}x_k$$

Where F_i: estimate of i^{th} factor, w_i: weight or factor score coefficient, x_i: i^{th} variable and k: number of variables

The coefficients are selected such that

- the first factor explains largest portion of the total variation
- the second factor accounts for the most of the residual variance, etc.

FACTOR ANALYSIS

- Helps to understand the variability in large data sets with inter correlated variables using a smaller number of uncorrelated factors.
- Explaining variability of a set of n variables using m factors where $m < n$
- The emphasis is on the identification of underlying factors that might explain the dimensions associated with large data

Objectives

- Reduces the complexity of a large set of variables by summarizing them in a smaller set of components or factors
- Tries to improve the interpretation of complex data through logical factors

FACTOR ANALYSIS

Steps

- Prepare correlation matrix
- Extract a set of factors using correlation matrix
- Determine the number of factors
- Rotate factors to increase interpretability
- Interpret results

FACTOR ANALYSIS

Example: Suppose a researcher wants to determine the underlying benefits consumers seek from the purchase of a toothpaste. A sample of 30 respondents was interviewed. The respondents were asked to indicate their degree of agreement with the following statements using a 7 point scale (1: strongly disagree, 7: strongly agree)

1. It is important to buy a toothpaste that prevents cavities
2. I like a toothpaste that gives shiny teeth
3. A toothpaste should strengthen your gums
4. I prefer toothpaste that freshens breath
5. Prevention of tooth decay is not an important benefit offered by a toothpaste
6. The most important consideration in buying a toothpaste is attractive teeth

FACTOR ANALYSIS

Step 1: Normalize the data

z transform:

Transformed data = (Data – Mean) / SD

Reading ghe file to R

```
>mydata = mydata[,2:7]
```

Transforming the variables

```
>myzdata = scale(mydata)
```

FACTOR ANALYSIS

Step 2: Check for Correlation

• Variables must be correlated for data reduction

> cor(myzdata)

Correlation Matrix

		x1	x2	x3	x4	x5	x6
Correlation	x1	1.000	-.053	.873	-.086	-.858	.004
	x2	-.053	1.000	-.155	.572	.020	.640
	x3	.873	-.155	1.000	-.248	-.778	-.018
	x4	-.086	.572	-.248	1.000	-.007	.640
	x5	-.858	.020	-.778	-.007	1.000	-.136
	x6	.004	.640	-.018	.640	-.136	1.000

High correlation between x1, x3 & x5

Good correlation between x2, x4 & x6

FACTOR ANALYSIS

Step 3: Check for Sampling (factor) adequacy

```
>library(psych)
> KMO(myzdata)
```

Statistics	Value	Criteria
Kaiser, Meyer, Olkin (KMO)	0.66	> 0.5

FACTOR ANALYSIS

Step 4: Method used: Principle Component Analysis

```
> mymodel = princomp(myzdata)
>summary(mymodel)
```

FACTOR ANALYSIS

Step 4: Method used: Principle Component Analysis

Used to identify minimum number of factors accounting for maximum variance in the data

Eigen Values: Amount of variance attributed to a component

Total Variance = 6 (Sum of all Eigen values)

Prop. variance for PC1= Eigen value of PC1 / Total Variance (2.731/6 = 0.455)

Component	SD	Variance	Proportion of Variance	Cumulative Proportion of Variance
PC 1	1.653	2.732	0.455	0.455
PC 2	1.489	2.217	0.369	0.825
PC 3	0.665	0.442	0.074	0.899
PC 4	0.584	0.341	0.057	0.955
PC 5	0.427	0.182	0.030	0.986
PC 6	0.292	0.085	0.014	1.000
Total		6.000		

FACTOR ANALYSIS

Step 4: Determine the number of Components

1. Based on Eigen Values: Only factors with Eigen value > 1.0 are selected

2. Based on cumulative % variance: Factors extracted should account for at least 65 % of variance

Component	SD	Variance	Proportion of Variance	Cumulative Proportion of Variance
PC 1	1.653	2.732	0.455	0.455
PC 2	1.489	2.217	0.369	0.825
PC 3	0.665	0.442	0.074	0.899
PC 4	0.584	0.341	0.057	0.955
PC 5	0.427	0.182	0.030	0.986
PC 6	0.292	0.085	0.014	1.000
Total		6.000		

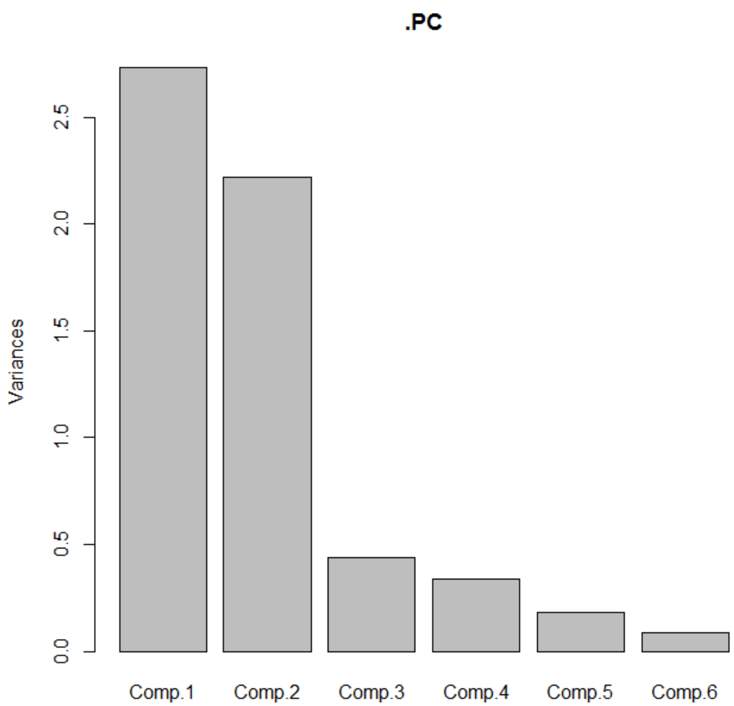
Number of factors selected : 2

FACTOR ANALYSIS

Step 4: Determine the number of Factors

```
>plot(mymodel)
```

3. Based on Scree plot: Plot of the eigen values against the number of factors in order of extraction. The number of factors is identified based on slope change of scree plot



Number of factors selected : 2

Step 5: Calculate Factor Scores– Eigen Vectors

>loadings(mymodel)

$$F_i = w_{i1}x_1 + w_{i2}x_2 + w_{i3}x_3 + ---- + w_{ik}x_k$$

	Component	
	1	2
x1	0.562	-0.170
x2	-0.182	-0.534
x3	0.566	-0.088
x4	-0.207	-0.530
x5	-0.526	0.236
x6	-0.107	-0.585

Step 5: Interpret Components – Eigen Vectors

	Component	
	1	2
x1	0.562	-0.170
x2	-0.182	-0.534
x3	0.566	-0.088
x4	-0.207	-0.530
x5	-0.526	0.236
x6	-0.107	-0.585

Component 1 is correlated with x1, x3 & x5

Component 2 is correlated with x2, x4 & x6

FACTOR ANALYSIS

Step 5: Interpret Components

	Component	
	1	2
Prevention of Cavities	0.562	-0.170
x2	-0.182	-0.534
Strong Gum	0.566	-0.088
x4	-0.207	-0.530
Non Prevention of Tooth Decay	-0.526	0.236
x6	-0.107	-0.585

Interpretation

Component 1 represents the health related benefits

Step 5: Interpret Components

	Component	
	1	2
Prevention of Cavities	0.562	-0.170
Shiny Teeth	-0.182	-0.534
Strong Gum	0.566	-0.088
Fresh Breath	-0.207	-0.530
Non Prevention of Tooth Decay	-0.526	0.236
Attractive Teeth	-0.107	-0.585

Interpretation

Component 2 represents the social related benefits

Step 6 : Varimax Rotation

Shows better relationship between variables and components

```
>library(psych)

>library(GPArotation)

>mymodel = principal(mydata, nfactors = 2, rotate = “varimax”)

<mymodel
```

	Component	
	1	2
x1	0.96	-0.03
x2	-0.05	0.85
x3	0.93	-0.15
x4	-0.09	0.85
x5	-0.93	-0.08
x6	0.09	0.88

Step 6: Reduced Data Set

```
>pc = mymodel$scores
>cbind(pc[,1], pc[,2])
```

Respondent	PC1	PC2	Respondent	PC1	PC2
1	-1.953	-0.071	16	-1.412	0.135
2	1.676	0.985	17	-1.261	0.610
3	-2.430	0.658	18	-2.504	-0.237
4	0.091	-1.697	19	1.298	1.397
5	1.515	2.724	20	1.278	-1.742
6	-1.670	0.015	21	1.449	1.791
7	-1.062	1.154	22	-0.978	-0.245
8	-2.088	-0.540	23	1.411	0.822
9	1.290	1.354	24	0.928	-2.680
10	2.796	-1.632	25	-1.431	-0.029
11	-2.040	0.389	26	1.079	-2.205
12	1.668	0.942	27	-1.470	0.106
13	-2.438	0.615	28	1.588	-1.216
14	0.425	-1.997	29	0.803	-3.270
15	1.651	1.880	30	1.790	1.987

FACTOR ANALYSIS

Exercise 1: Data on Customer satisfaction survey conducted by IT company is given below. Each customer is asked to were asked to indicate their degree of agreement with the following statements using a 7 point scale (1: strongly disagree, 7: strongly agree) . Can you reduce the 14 variables into less number of factors?

CLUSTER ANALYSIS

CLUSTER ANALYSIS

A technique used to classify objects or cases into relatively homogeneous groups called clusters

Cluster

A collection of data objects similar to one another within the same cluster and dissimilar to the objects in other clusters

Cluster analysis

A procedure for grouping a set of data objects into clusters

CLUSTER ANALYSIS

- A technique used to classify objects or cases into relatively homogeneous groups called clusters

Example: A survey was done to study the consumers attitude towards shopping. The consumers need to be clustered based on their attitude towards shopping. The respondents were asked to express their degree of agreement with the following statements on a 7 point scale (1: strongly disagree, 7: strongly agree).

x1: Shopping is fun

x2: Shopping is bad for your budget

x3: I combine shopping with eating out

x4: I try to get the best buys when shopping

x5: I don’t care about shopping

x6: You can save a lot os money by comparing prices

CLUSTER ANALYSIS

Step 1: Choose Type of clustering - Agglomerative Clustering

- Hierarchical Clustering – characterized by development of a hierarchy or tree like structure
- Starts with each object or record as separate clusters
- Clusters are formed by grouping objects in to bigger and bigger clusters until all objects are in one cluster.
- The objects grouped based on linkage measure

CLUSTER ANALYSIS

Types of Linkage

1. Single Linkage:

   Based on minimum distance

   The first two objects clustered are those having minimum distance between them

2. Complete Linkage:

   Based on maximum distance

   The distance between two clusters is calculated as the distance between two furthest points

3. Average Linkage:

   Based on average distance

   The distance between two clusters is defined as the average of the distance between all pairs of points

   Preferred method

Step 2: Choose Method

Variance method:

Generates clusters with minimum within cluster variance

Uses Ward's Procedure

Ward's Procedure

For each cluster means for all the variables are computed

For each object or record, the squared Euclidean distance to the cluster mean is computed

CLUSTER ANALYSIS

R Code

Read data to mydata and compute distance

```
> distance = dist(mydata, method = “eucldean”)
```

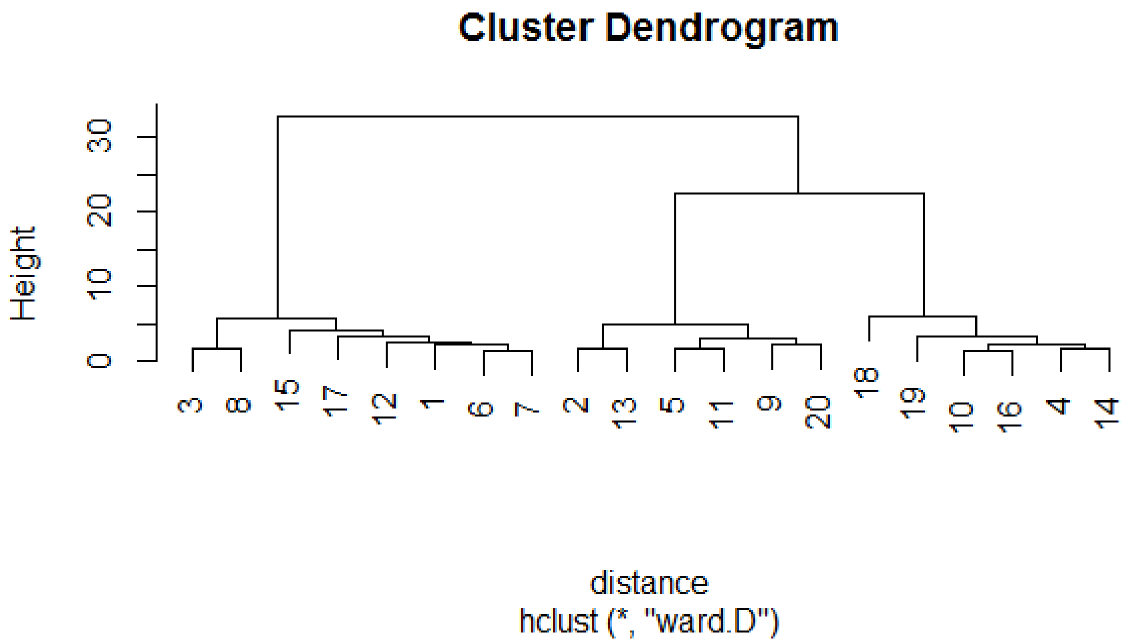
Generate Clusters

```
>mymodel = hclust(distance, method = “ward”)
```

Plot Dendogram

```
>plot(mymodel)
```

Decide on number of clusters: Dendrogram



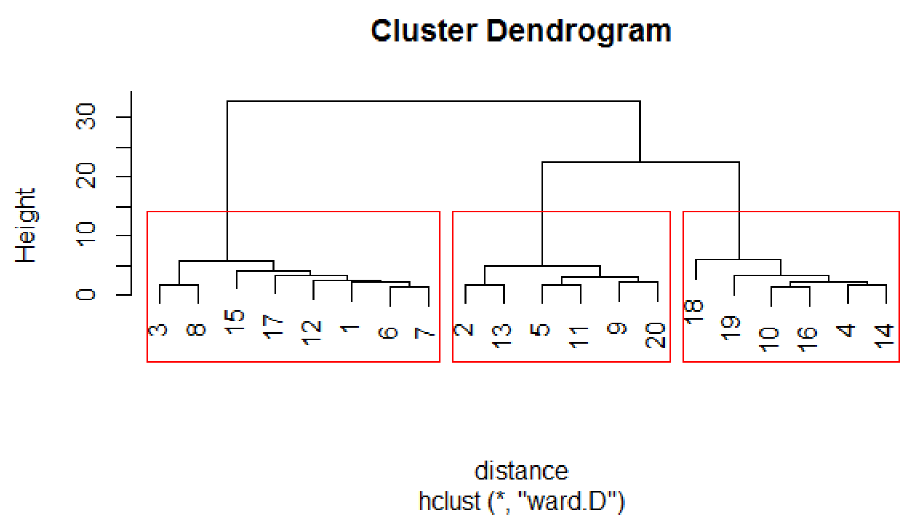
CLUSTER ANALYSIS

Decide on number of clusters: Dendrogram

Stages is given in x axis and distance in y axis

When one move from 3 cluster to 2 cluster the distance increases drastically. So 3 cluster may be appropriate

```
>groups = cutree(mymodel, k = 3)
> rect.hclust(mymodel, k = 3, border = "red")
```



Identification of cluster membership for each record

```
>mynewmodel = kmeans(mydata,3)
>cluster = mynewmodel$cluster
>output = cbind(mydata, cluster)
>write,csv(output, "E:/ISI_Mumbai/output.csv“)
```

Cluster membership

Indicates each record or case falls in which cluster based on number of clusters

	x1	x2	x3	x4	x5	x6	cluster
1	6	4	7	3	2	3	3
2	2	3	1	4	5	4	2
3	7	2	6	4	1	3	3
4	4	6	4	5	3	6	1
5	1	3	2	2	6	4	2
6	6	4	6	3	3	4	3
7	5	3	6	3	3	4	3
8	7	3	7	4	1	4	3
9	2	4	3	3	6	3	2
10	3	5	3	6	4	6	1
11	1	3	2	3	5	3	2
12	5	4	5	4	2	4	3
13	2	2	1	5	4	4	2
14	4	6	4	6	4	7	1
15	6	5	4	2	1	4	3
16	3	5	4	6	4	7	1
17	4	4	7	2	2	5	3
18	3	7	2	6	4	3	1
19	4	6	3	7	2	7	1
20	2	3	2	4	7	2	2

CLUSTER ANALYSIS

Cluster Profile

```
> aggregate(mydata, by = list(cluster), FUN = mean)
```

Variables	Cluster Means		
	1	2	3
x1 (shopping is fun)	3.50	1.67	5.75
x2 (shopping upsets my budget)	5.83	3.00	3.63
x3 (I combine shopping with eating out)	3.33	1.83	6.00
x4 (I try to get best buys when shopping)	6.00	3.50	3.13
x5 (I don't care about shopping)	3.50	5.50	1.88
X6 (save a lot by comparing prices	6.00	3.33	3.88

Cluster 1: High on x2 x4 & x6
Concerned about spending money (Economical)

Cluster 2: Low on x1 & x3 but High on x5
Careless & no fun in shopping (apathetic)

Cluster 3: High on x1 & x3 but low on x5
Fun loving and concerned

CLUSTER ANALYSIS

Exercise 1: Data on Customer satisfaction survey conducted by IT company is given below. Each customer is asked to were asked to indicate their degree of agreement with the following statements using a 7 point scale (1: strongly disagree, 7: strongly agree) . Can you group the customers into meaningful groups?

NAÏVE BAYES CLASSIFIER

NAÏVE BAYES CLASSIFIER

- A graph together with an associated set of probability tables
- The nodes of the graph represent variables and the arcs represent the relationship between the variables
- Used to model the dependencies between all the variables in the data
- Model the joint probability distribution of the variables
- Used to predict the probability that the value of the output variable will fall in an interval for a given set of values of input or predictor variables

NAÏVE BAYES CLASSIFIER

Review Duration	Code Coverage	Defect Density
1 to 2 hours	Medium	> 3.3
1 to 2 hours	Medium	> 3.3
2 to 3 hours	High	< 3.3
1 to 2 hours	Medium	> 3.3
1 to 2 hours	High	> 3.3
2 to 3 hours	High	< 3.3
2 to 3 hours	Medium	< 3.3
1 to 2 hours	Medium	> 3.3
2 to 3 hours	High	< 3.3
2 to 3 hours	Medium	> 3.3

$$X : \mathrm{Re}view \;\; duration = 2\ to\ 3hrs\ \&\ code\ \mathrm{cov}erage = medium$$

$$P(defect\ density < 3.3) = \frac{4}{10} = 0.4$$

$$P(defect\ density \geq 3.3) = \frac{6}{10} = 0.6$$

NAÏVE BAYES CLASSIFIER

Review Duration	Code Coverage	Defect Density
1 to 2 hours	Medium	> 3.3
1 to 2 hours	Medium	> 3.3
2 to 3 hours	High	< 3.3
1 to 2 hours	Medium	> 3.3
1 to 2 hours	High	> 3.3
2 to 3 hours	High	< 3.3
2 to 3 hours	Medium	< 3.3
1 to 2 hours	Medium	> 3.3
2 to 3 hours	High	< 3.3
2 to 3 hours	Medium	> 3.3

$$X : \text{Review} \;\; duration = 2 \; to \; 3hrs \;\&\; code \; \text{coverage} = medium$$

$$P(review \; duration = 2 \; to \; 3 \; hrs \; / \; defect \; density < 3.3) = \frac{4}{4} = 1$$

$$P(code \; \text{coverage} = medium \, / \; defect \; density < 3.3) = \frac{1}{4} = 0.25$$

$$P(X \;\; / \, defect \;\; density < 3.3) = 1 \times 0.25 = 0.25$$

$$P(X \;\; / \, defect \;\; density < 3.3) \times P(defect \;\; density < 3.3) = 0.25 \times 0.4 = 0.1$$

NAÏVE BAYES CLASSIFIER

Review Duration	Code Coverage	Defect Density
1 to 2 hours	Medium	> 3.3
1 to 2 hours	Medium	> 3.3
2 to 3 hours	High	< 3.3
1 to 2 hours	Medium	> 3.3
1 to 2 hours	High	> 3.3
2 to 3 hours	High	< 3.3
2 to 3 hours	Medium	< 3.3
1 to 2 hours	Medium	> 3.3
2 to 3 hours	High	< 3.3
2 to 3 hours	Medium	> 3.3

$$X : \text{Review } \; duration = 2 \; to \; 3hrs \; \& \; code \; \text{coverage} = medium$$

$$P(review \; duration = 2 \; to \; 3 \; hrs \, / \, defect \; density \geq 3.3) = \frac{1}{6} = 0.17$$

$$P(code \; \text{coverage} = medium \, / \, defect \; density \geq 3.3) = \frac{5}{6} = 0.83$$

$$P(X \; / \, defect \; density \geq 3.3) = 0.17 \times 0..83 = 0.1389$$

$$P(X \; / \, defect \; density \geq 3.3) \times P(defect \; density \geq 3.3) = 0.1389 \times 0.6 = 0.0833$$

NAÏVE BAYES CLASSIFIER

Review Duration	Code Coverage	Defect Density
1 to 2 hours	Medium	> 3.3
1 to 2 hours	Medium	> 3.3
2 to 3 hours	High	< 3.3
1 to 2 hours	Medium	> 3.3
1 to 2 hours	High	> 3.3
2 to 3 hours	High	< 3.3
2 to 3 hours	Medium	< 3.3
1 to 2 hours	Medium	> 3.3
2 to 3 hours	High	< 3.3
2 to 3 hours	Medium	> 3.3

$$X : \mathrm{Re}view \quad duration = 2 \ to \ 3hrs \ \& \ code \ \mathrm{coverage} = medium$$

$$P(defect \ density < 3.3 / X) = \frac{0.1}{0.1 + 0.0833} = 0.545 = 54.5\%$$

$$P(defect \ density \geq 3.3 / X) = \frac{0.0833}{0.1 + 0.0833} = 0.454 = 45.4\%$$

NAÏVE BAYES CLASSIFIER

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

Example: Develop a model to predict the iris plant class based on sepal length, sepal width, petal length and petal width. The data is given in Iris.csv file. Validate the model with Iris_test.csv data?

NAÏVE BAYES CLASSIFIER

Example: Develop a model to predict the iris plant class based on sepal length, sepal width, petal length and petal width using Naïve Bayes classifier. The data is given in Iris.csv file. Validate the model with Iris_test.csv data?

1. Read file

```
> mydata = Iris
```

2. Call library e1071

```
> libray(e1071)
```

3. Develop Model

```
> model = naiveBayes(mydata[,1:4], mydata[,5])
> model
```

NAÏVE BAYES CLASSIFIER

Example: Develop a model to predict the iris plant class based on sepal length, sepal width, petal length and petal width using Naïve Bayes classifier. The data is given in Iris.csv file. Validate the model with Iris_test.csv data?

4. Compute Predicted values

```
> pred = predict(model, mydata[,1:4])
> pred
```

5. Model evaluation (Actual Vs Predicted)

```
> mytable = table(pred, mydata[,5])
> mytable
```

Predicted	Actual		
	Iris-setosa	Iris-versicolor	Iris-virginica
Iris-setosa	50	0	0
Iris-versicolor	0	47	3
Iris-virginica	0	3	47

NAÏVE BAYES CLASSIFIER

Example: Develop a model to predict the iris plant class based on sepal length, sepal width, petal length and petal width using Naïve Bayes classifier. The data is given in Iris.csv file. Validate the model with Iris_test.csv data?

6. Reading test data file

```
> mytestdata = Iris_test
```

7. Predicting output for test data

```
> predtest = predict(model, mytestdata[,1:4])
> predtest
```

NAÏVE BAYES CLASSIFIER

Example: Develop a model to predict the iris plant class based on sepal length, sepal width, petal length and petal width using Naïve Bayes classifier. The data is given in Iris.csv file. Validate the model with Iris_test.csv data?

8. Model evaluation using test data(Actual Vs Predicted)

```
> mytesttable = table(predtest,mytestdata[,5])
> mytesttable
```

Predicted	Actual		
	Iris-setosa	Iris-versicolor	Iris-virginica
Iris-setosa	49	0	0
Iris-versicolor	0	14	0
Iris-virginica	0	1	2

FORECASTING

INTRODUCTION

Time Series:

A collection of observations or data made sequentially in time.

A dataset consisting of observations arranged in chronological order

A sequence of observations over time

Forecast:

An estimate of the future value of some variable

Example:

The number of 2 wheeler sales in Bangalore during next month

The average volume of an airline passengers in the next quarter

INTRODUCTION

Time Series Plot:

The graphical representation of time series data by taking time on x axis & data on y axis.

A plot of data over time

Example

The demand for a commodity E15 for last 20 months is given below. Draw the time series plot

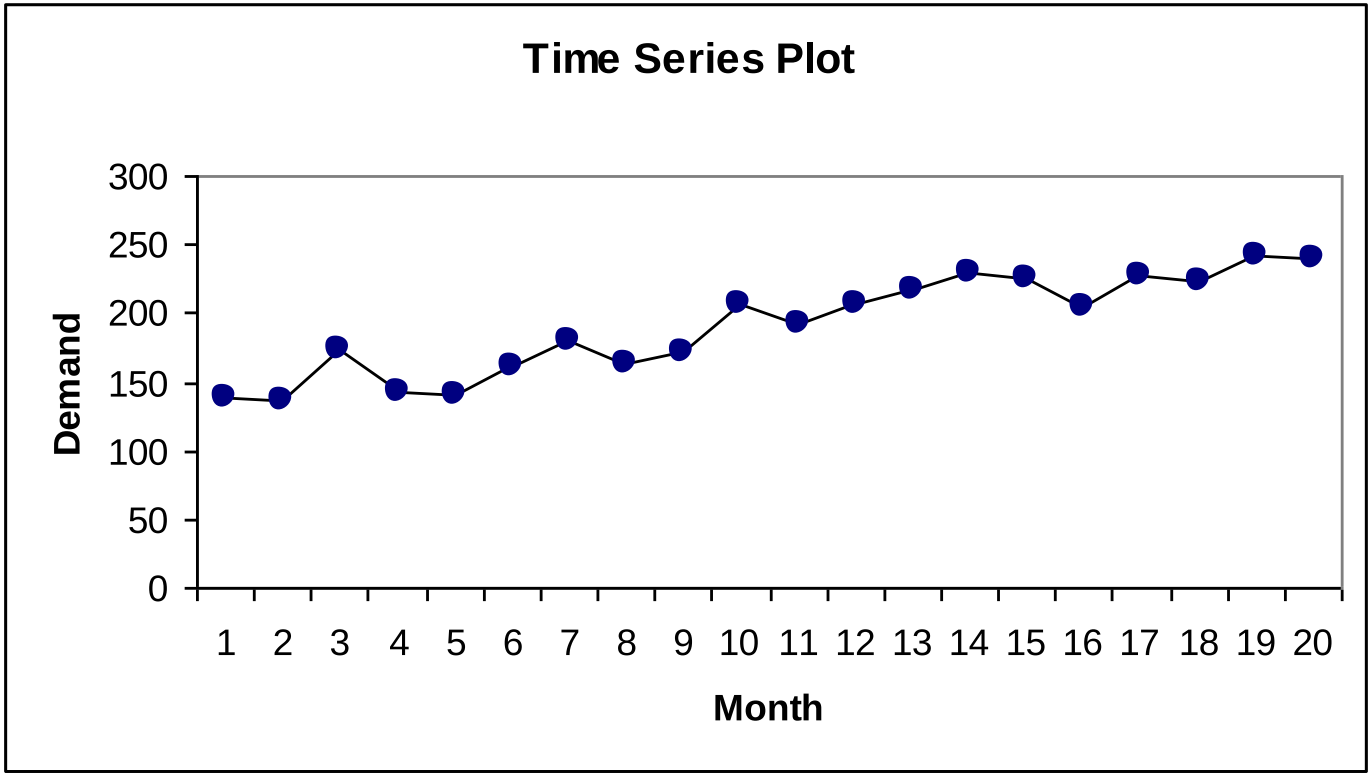
Month	Demand	Month	Demand
1	139	11	193
2	137	12	207
3	174	13	218
4	142	14	229
5	141	15	225
6	162	16	204
7	180	17	227
8	164	18	223
9	171	19	242
10	206	20	239

INTRODUCTION

Time Series Plot:

Example

Time series plot of the demand for a commodity E15



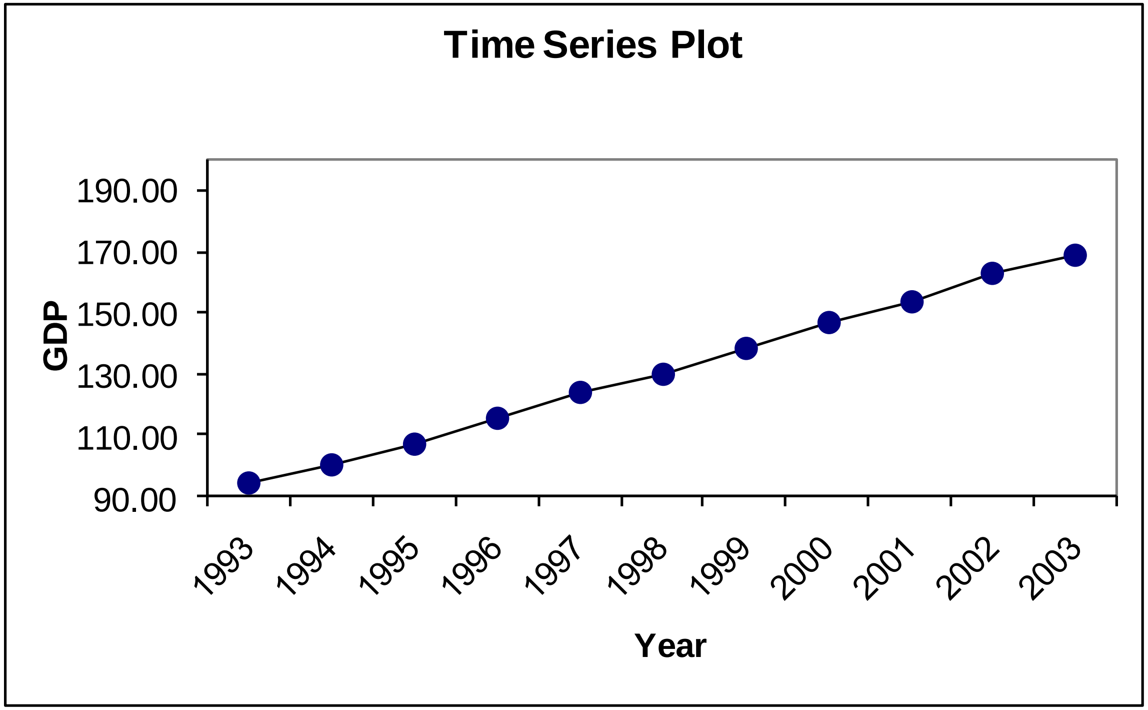
INTRODUCTION

Trend:

A long term increase or decrease in the data

Example: The data on Yearly average of Indian GDP during 1993 to 2005.

Year	GDP
1993	94.43
1994	100.00
1995	107.25
1996	115.13
1997	124.16
1998	130.11
1999	138.57
2000	146.97
2001	153.40
2002	162.28
2003	168.73



INTRODUCTION

Seasonal Pattern:

The time series data exhibiting rises and falls influenced by seasonal factors

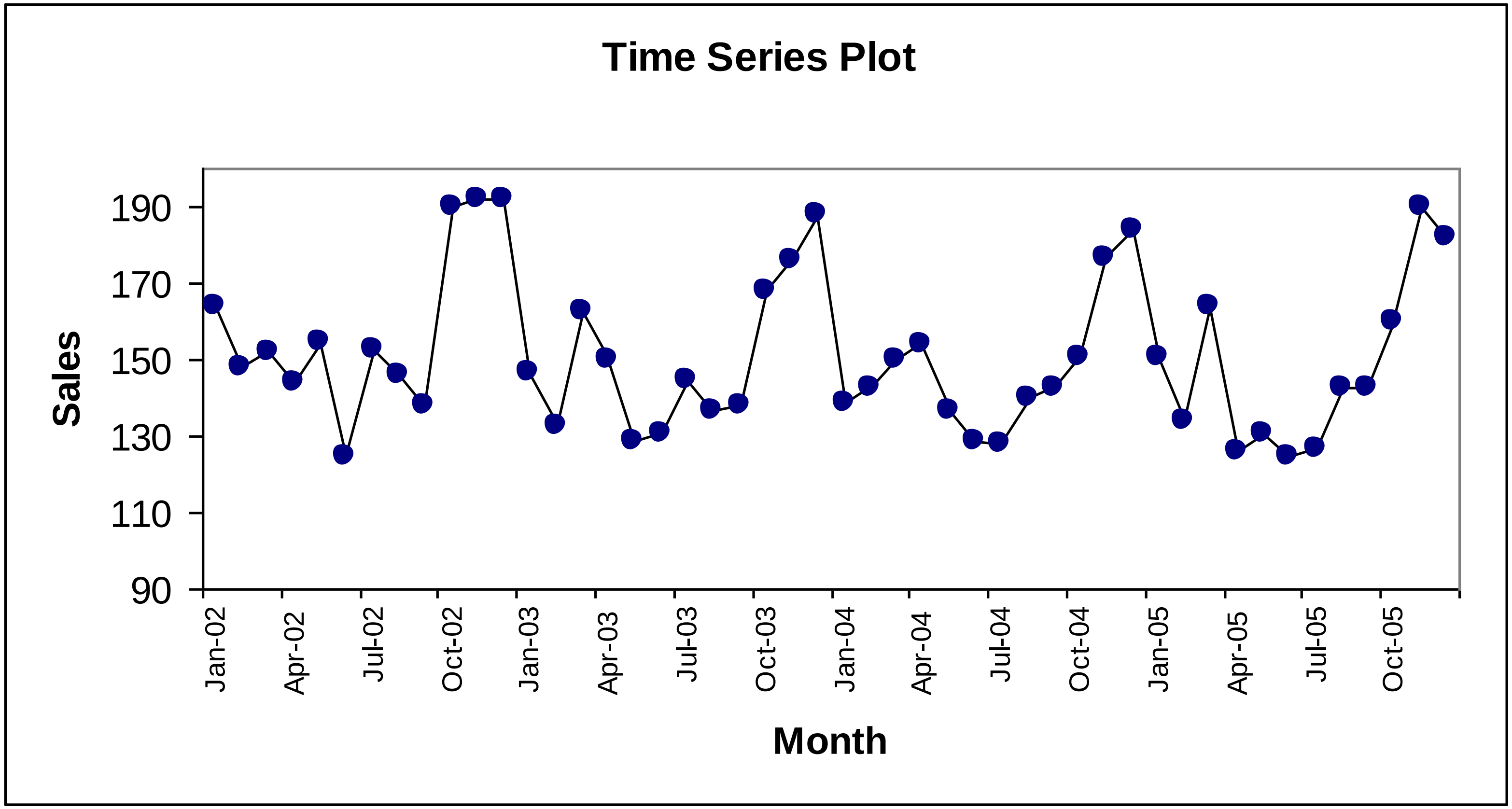
Example: The data on monthly sales of a branded jackets

Month	Sales	Month	Sales	Month	Sales	Month	Sales
Jan-02	164	Jan-03	147	Jan-04	139	Jan-05	151
Feb-02	148	Feb-03	133	Feb-04	143	Feb-05	134
Mar-02	152	Mar-03	163	Mar-04	150	Mar-05	164
Apr-02	144	Apr-03	150	Apr-04	154	Apr-05	126
May-02	155	May-03	129	May-04	137	May-05	131
Jun-02	125	Jun-03	131	Jun-04	129	Jun-05	125
Jul-02	153	Jul-03	145	Jul-04	128	Jul-05	127
Aug-02	146	Aug-03	137	Aug-04	140	Aug-05	143
Sep-02	138	Sep-03	138	Sep-04	143	Sep-05	143
Oct-02	190	Oct-03	168	Oct-04	151	Oct-05	160
Nov-02	192	Nov-03	176	Nov-04	177	Nov-05	190
Dec-02	192	Dec-03	188	Dec-04	184	Dec-05	182

INTRODUCTION

Seasonal Pattern:

The time series data exhibiting rises and falls influenced by seasonal factors

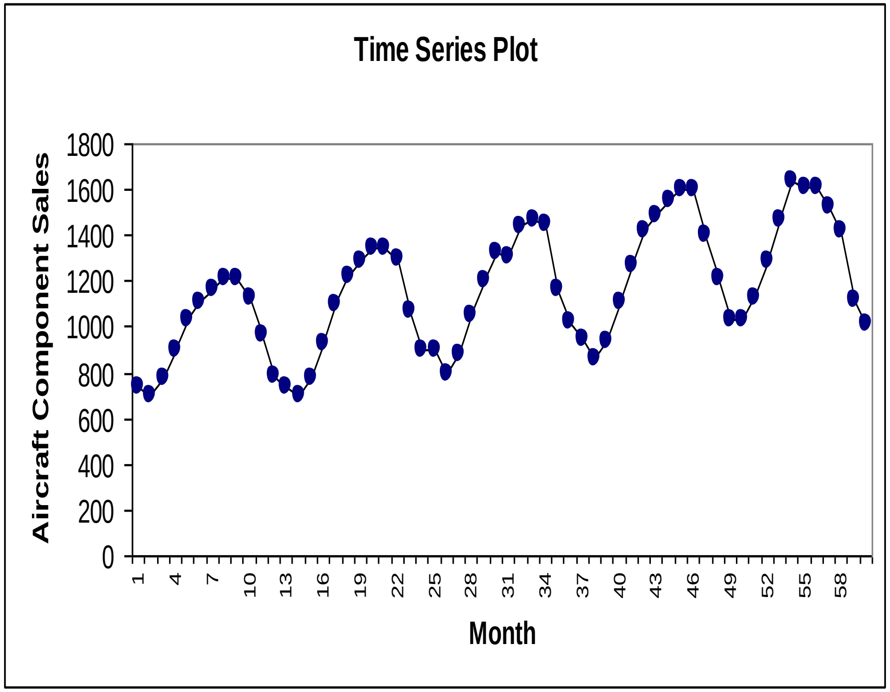


INTRODUCTION

The time series data may include a combination of trend and seasonal patterns

Example: The data on monthly sales of an aircraft component is given below:

Month	Sales	Month	Sales	Month	Sales
1	742	21	1341	41	1274
2	697	22	1296	42	1422
3	776	23	1066	43	1486
4	898	24	901	44	1555
5	1030	25	896	45	1604
6	1107	26	793	46	1600
7	1165	27	885	47	1403
8	1216	28	1055	48	1209
9	1208	29	1204	49	1030
10	1131	30	1326	50	1032
11	971	31	1303	51	1126
12	783	32	1436	52	1285
13	741	33	1473	53	1468
14	700	34	1453	54	1637
15	774	35	1170	55	1611
16	932	36	1023	56	1608
17	1099	37	951	57	1528
18	1223	38	861	58	1420
19	1290	39	938	59	1119
20	1349	40	1109	60	1013



INTRODUCTION

Stationary Series: A series from trend and seasonal patterns.

A series exhibits only random fluctuations around mean

Example : The data on daily shipments is given in table below> Check whether the data is stationary

Day	Shipments	Day	Shipments
1	99	13	101
2	103	14	111
3	92	15	94
4	100	16	101
5	99	17	104
6	99	18	99
7	103	19	94
8	101	20	110
9	100	21	108
10	100	22	102
11	102	23	100
12	101	24	98

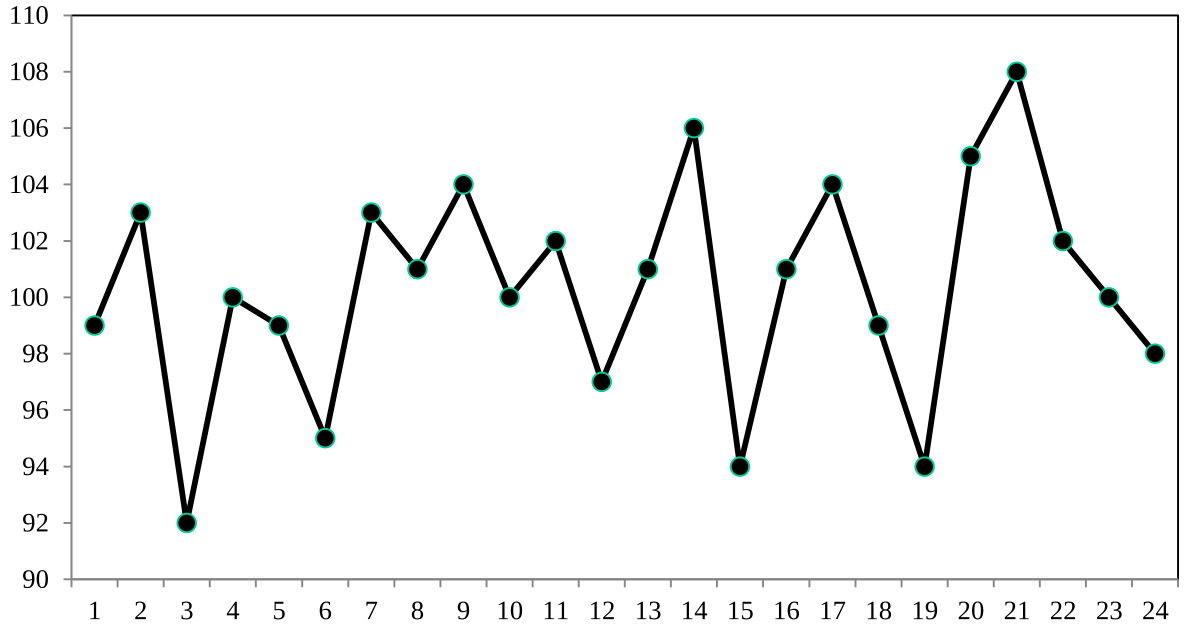
```
mydata=ts(ship[,"Shipments"])
plot(mydata)
```

INTRODUCTION

Stationary Series: A series from trend and seasonal patterns.

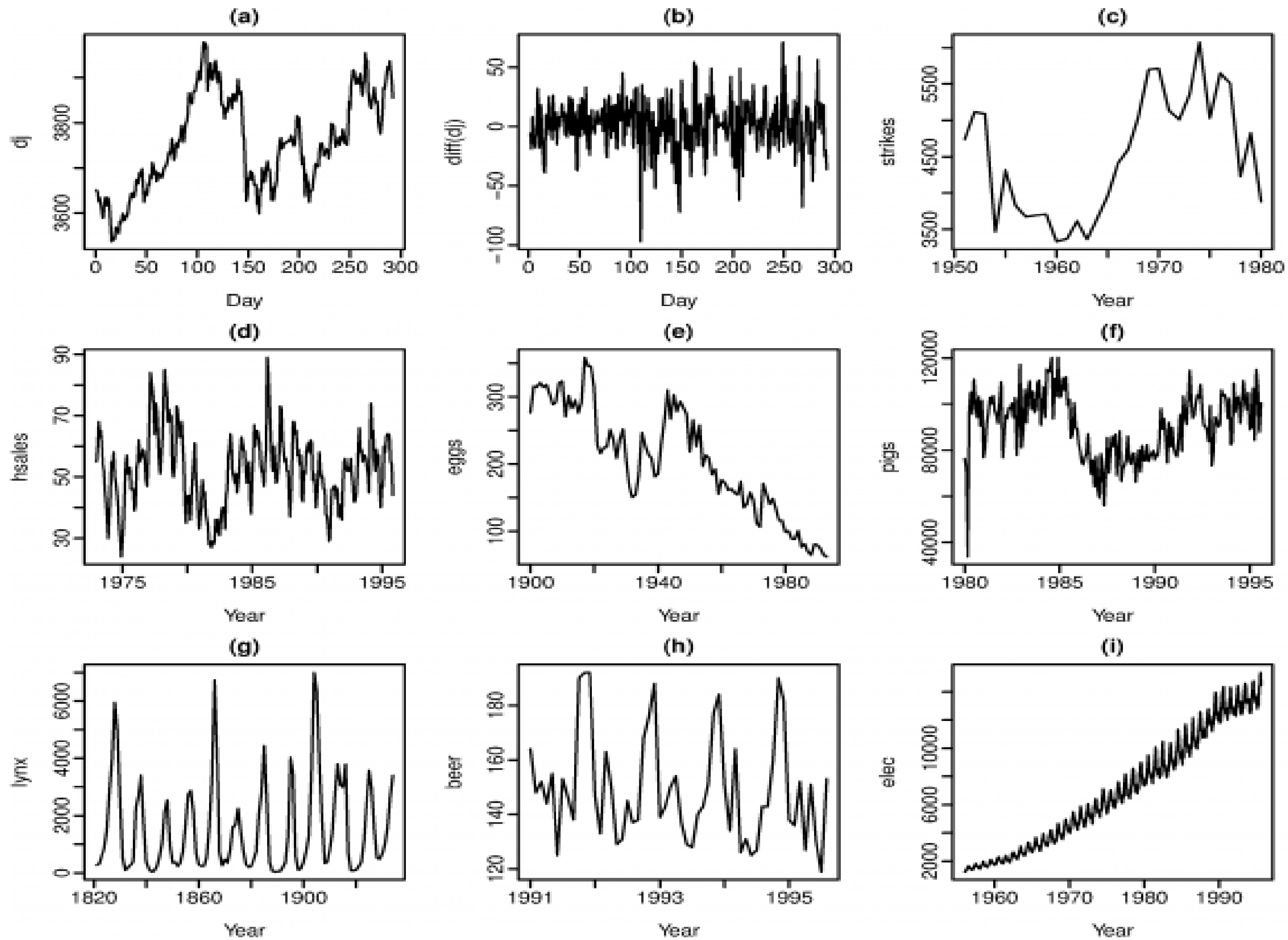
A series exhibits only random fluctuations around mean

Example : The data on daily shipments is given in table below. Check whether the data is stationary



INTRODUCTION

Stationary Series: A series from trend and seasonal patterns. A series exhibits only random fluctuations around mean



INTRODUCTION

Test for Stationary: Unit root test

Augmented Dickey Fuller Test (ADF) :

If the test statistic value is smaller than the relevant critical value (generally 5%), then the data is stationary. The Null hypothesis of ADF test is data is non-stationary. A small p-value suggest data is stationary.

Kwiatkowski-Phillips-Schmidt-Shin Test (KPSS) :

Another test for stationary. The Null hypothesis of ADF test is data is stationary. A large p-value suggest data is stationary.

Example : Check whether the data on daily shipments is stationary

INTRODUCTION

Test for Stationary: Unit root test in R

- adf.test(mydata,alternative="stationary")

- Augmented Dickey-Fuller Test
  data: mydata
  Dickey-Fuller = -3.2471, Lag order = 2, p-value = 0.09901 alternative hypothesis: stationary

- kpss.test(mydata)

- KPSS Test for Level Stationarity
  data: mydata
  KPSS Level = 0.1967, Truncation lag parameter = 1, p-value = 0.1

Warning message: In kpss.test(mydata) : p-value greater than printed p-value

- ndiffs(mydata)
- [1] 0

INTRODUCTION

Test for Stationary: Unit root test

Augmented Dickey Fuller Test (ADF) :

If the test statistic value is smaller than the relevant critical value (generally 5%), then the data is stationary

Exercise : Check whether the GDP data is stationary

Year	GDP
1993	94.43
1994	100.00
1995	107.25
1996	115.13
1997	124.16
1998	130.11
1999	138.57
2000	146.97
2001	153.40
2002	162.28
2003	168.73

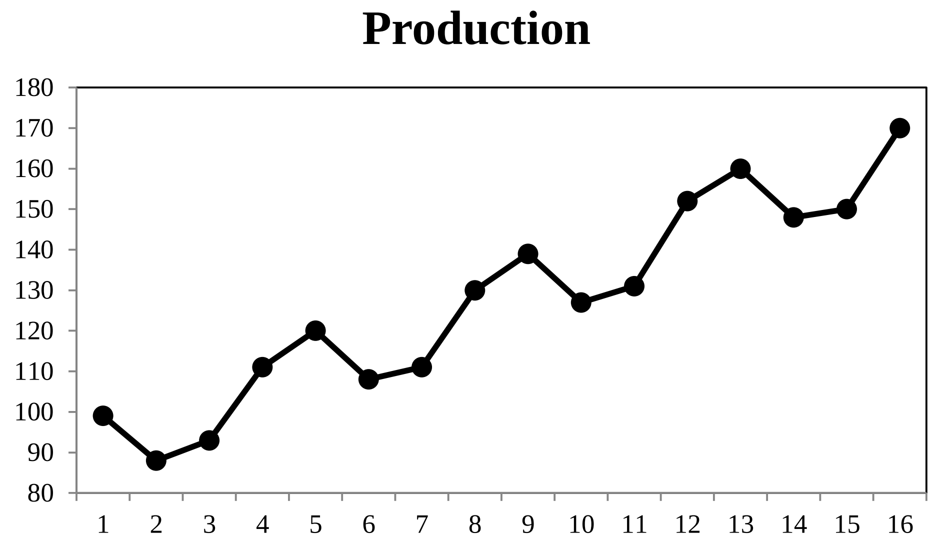
INTRODUCTION

Test for Stationary: Unit root test

Augmented Dickey Fuller Test (ADF) :

Exercise : Check whether the manganese production data is stationary

Period	Quarter	Production
1	1	99
2	2	88
3	3	93
4	4	111
5	1	120
6	2	108
7	3	111
8	4	130
9	1	139
10	2	127
11	3	131
12	4	152
13	1	160
14	2	148
15	3	150
16	4	170



INTRODUCTION

Differencing: A method for making data stationary

A differenced series is the series of difference between each observation Yt and the previous observation Yt-1

$$Yt' = Yt - Yt\text{-}1$$

A series with trend can be made stationary with 1st differencing

A series with seasonality can be made stationary with seasonal differencing

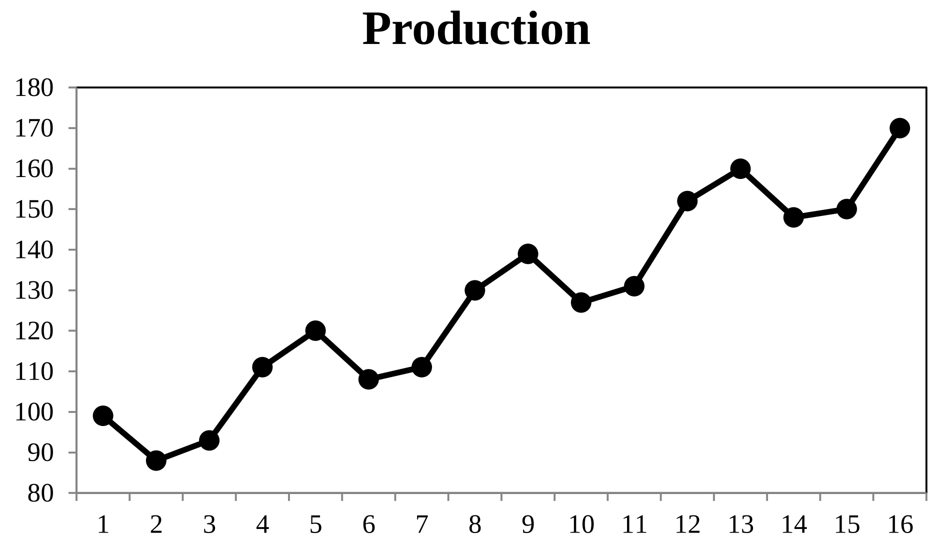
Example: Is it possible to make the GDP data given below stationary

INTRODUCTION

Differencing: Example

Is it possible to make the Manganese production data given below stationary

Period	Quarter	Production
1	1	99
2	2	88
3	3	93
4	4	111
5	1	120
6	2	108
7	3	111
8	4	130
9	1	139
10	2	127
11	3	131
12	4	152
13	1	160
14	2	148
15	3	150
16	4	170



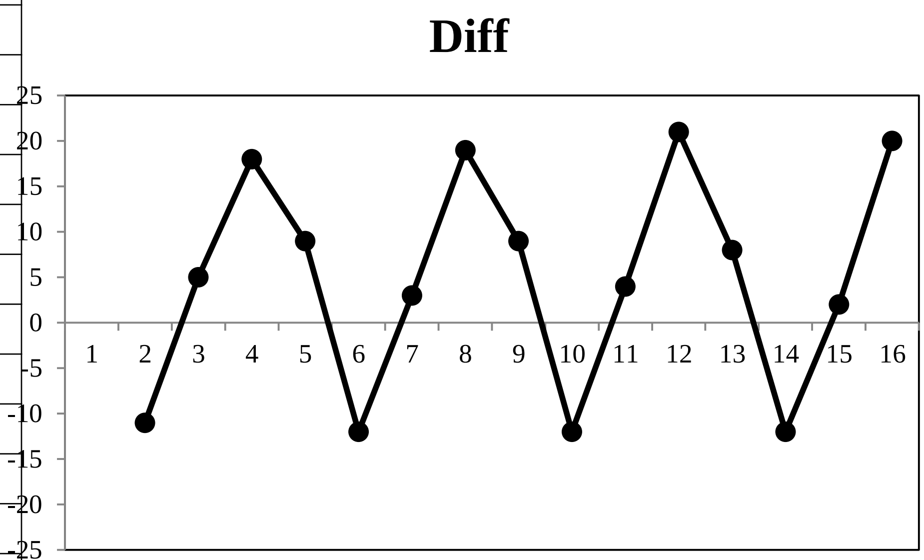
newdata=diff(mydata,1)

INTRODUCTION

Differencing: Example

Is it possible to make the Manganese production data given below stationary

Period	Quarter	Production	Diff
1	1	99	
2	2	88	-11
3	3	93	5
4	4	111	18
5	1	120	9
6	2	108	-12
7	3	111	3
8	4	130	19
9	1	139	9
10	2	127	-12
11	3	131	4
12	4	152	21
13	1	160	8
14	2	148	-12
15	3	150	2
16	4	170	20



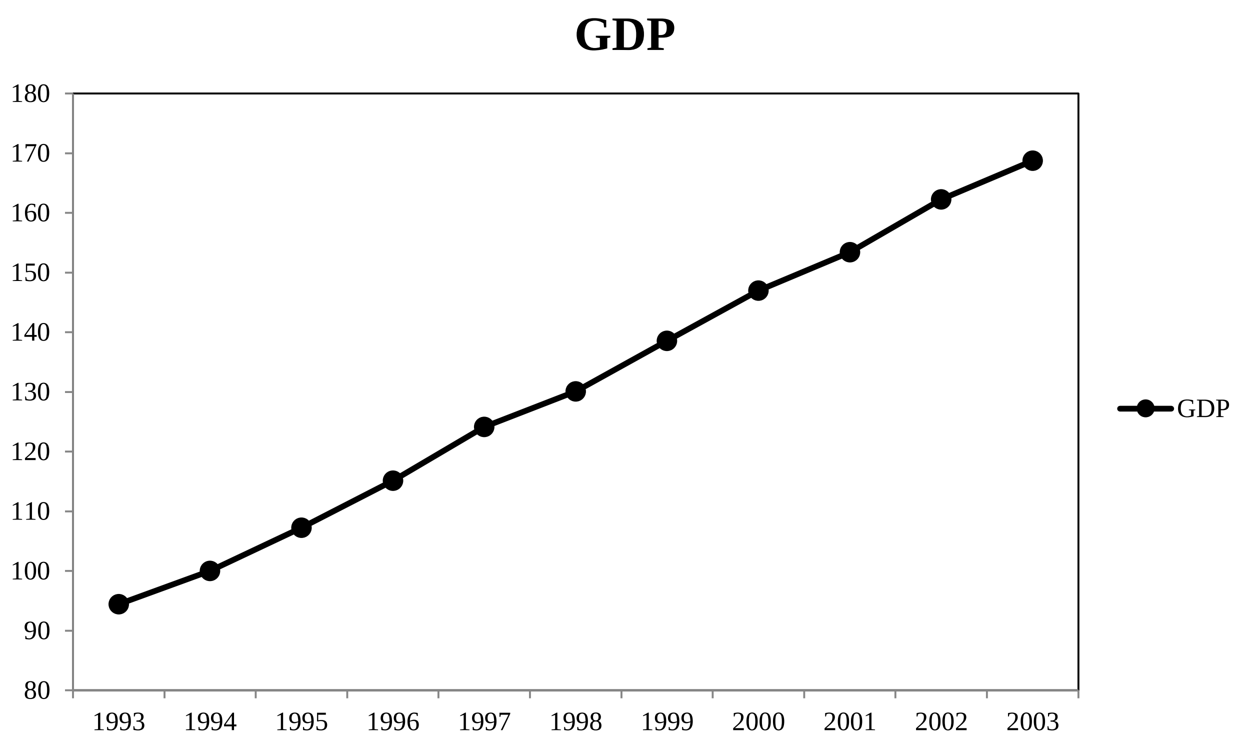
kpss.test(newdata)

INTRODUCTION

Differencing: Example

Is it possible to make the GDP data given below stationary

Year	GDP
1993	94.43
1994	100
1995	107.3
1996	115.1
1997	124.2
1998	130.1
1999	138.6
2000	147
2001	153.4
2002	162.3
2003	168.7

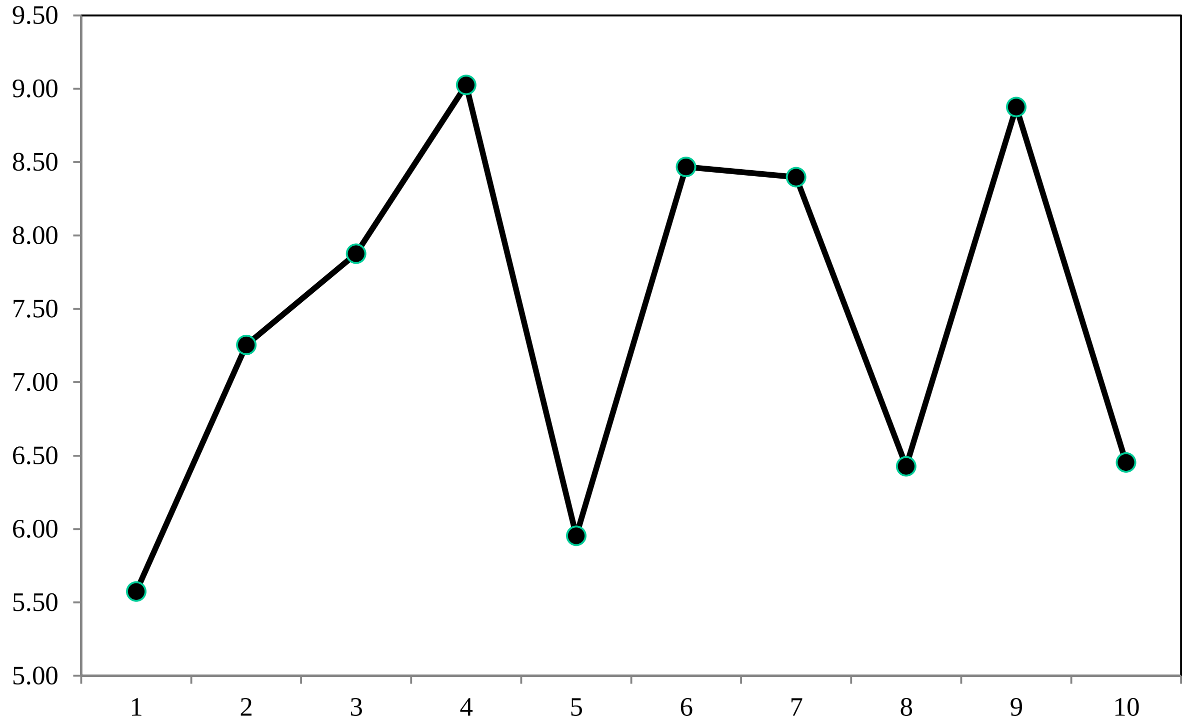


INTRODUCTION

Differencing: Example

Is it possible to make the GDP data given below stationary

Year	GDP	Diff
1993	94.43	
1994	100	5.57
1995	107.3	7.25
1996	115.1	7.88
1997	124.2	9.03
1998	130.1	5.95
1999	138.6	8.46
2000	147	8.4
2001	153.4	6.43
2002	162.3	8.88
2003	168.7	6.45



MODELING

General form of linear model

y is modeled in terms of x's

$$Y = a + b_1x_1 + b_2x_2 + - - - + b_kx_k$$

Step 1: Check Correlation between y and x's

y should be correlated with some of the x's

Time series model

Generally there will not be any x's

Hence patterns in y series is explored

y will be modeled in terms of previous values of y

$$y_t = a + b_1y_{t-1} + b_2y_{t-2} + - -$$

Step 1: Check correlation between y_t and y_{t-1}, etc

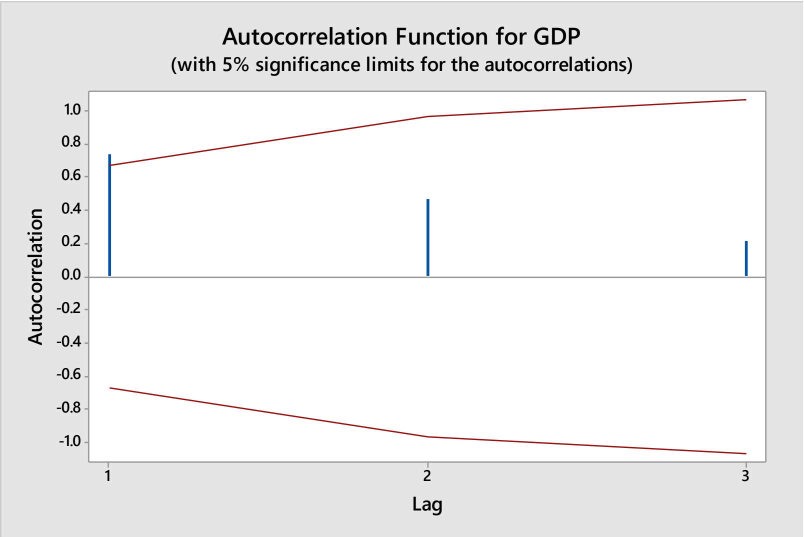
correlation between y and previous values of y are called **autocorrelation**

MODELING - ACF

Example: Check the auto correlation up to 3 lags in GDP data

Year	GDP(y_t)	y_{t-1}	y_{t-2}	y_{t-3}
1993	94.43			
1994	100	94.43		
1995	107.3	100	94.43	
1996	115.1	107.3	100	94.43
1997	124.2	115.1	107.3	100
1998	130.1	124.2	115.1	107.3
1999	138.6	130.1	124.2	115.1
2000	147	138.6	130.1	124.2
2001	153.4	147	138.6	130.1
2002	162.3	153.4	147	138.6
2003	168.7	162.3	153.4	147

Lag	variables	Auto Correlation
1	y_t vs y_{t-1}	0.7391
2	y_t vs y_{t-2}	0.4681
3	y_t vs y_{t-3}	0.2201



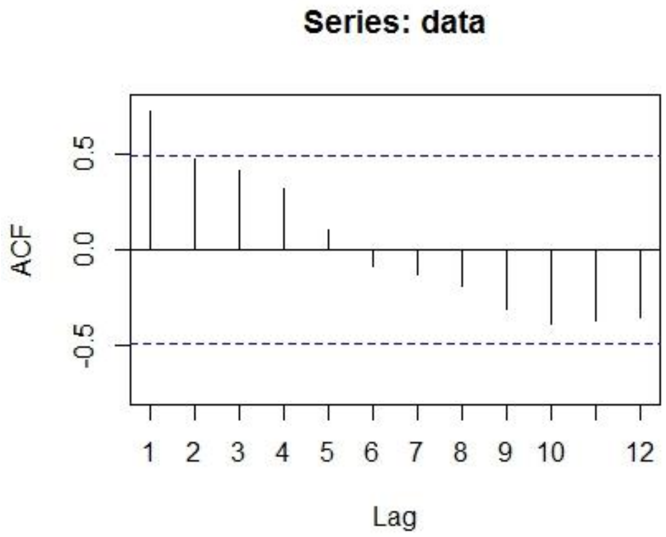
$$r_k = \frac{\sum_{i=1}^{n-k} (y_{k+i} - \bar{y})(y_i - \bar{y})}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

https://onlinecourses.science.psu.edu/stat510/node/62

MODELING

Example: Check the auto correlation up to 5 lags in Manganese Production data

Period	Quarter	Production
1	1	99
2	2	88
3	3	93
4	4	111
5	1	120
6	2	108
7	3	111
8	4	130
9	1	139
10	2	127
11	3	131
12	4	152
13	1	160
14	2	148
15	3	150
16	4	170



MODELING - PACF

- A partial correlation is a conditional correlation. It is the correlation between two variables under the assumption that we know and take into account the values of some other set of variables.

- For instance, consider a regression context in which y = response variable and x_1, x_2, and x_3 are predictor variables. The partial correlation between y and x_3 is the correlation between the variables determined taking into account how both y and x_3 are related to x_1 and x_2.

- In regression, this partial correlation could be found by correlating the residuals from two different regressions: (1) Regression in which we predict y from x_1 and x_2, (2) regression in which we predict x_3 from x_1 and x_2. Basically, we correlate the "parts" of y and x_3 that are not predicted by x_1 and x_2.

Some Useful Facts About PACF and ACF Patterns

1. Identification of an AR model is often best done with the PACF.
2. Identification of an MA model is often best done with the ACF rather than the PACF.

FORECAST ACCURACY MEASURES

Mean Absolute Error: MAE

Mean Square Error: MSE

Mean Absolute Percentage Error: MAPE

Weighted Mean Absolute Percentage Error: WMAPE

Example: The data on Yearly average of Indian GDP during 1993 to 2005. The predicted values using a suitable forecasting method is also given. Check the forecast accuracy using MAE

Year	GDP	Predicted
1993	94.43	91
1994	100.00	99.165
1995	107.25	107.329
1996	115.13	115.494
1997	124.16	123.659
1998	130.11	131.824
1999	138.57	139.989
2000	146.97	148.154
2001	153.40	156.319
2002	162.28	164.484
2003	168.73	172.649
2004	183.09	180.814
2005	195.74	188.979

FORECAST ACCURACY MEASURES

Mean Absolute Error : MAE

Step 1: Calculate Error : Error = Actual - Predicted

Step 2: Calculate absolute Error : Absolute Error = absolute (Actual – Predicted)

Step 3: Calculate MAE : MAE = Average of Absolute Error

Year	GDP	Predicted	Error	Absolute (Error)
1993	94.43	91	3.43	3.42589
1994	100.00	99.165	0.83	0.83500
1995	107.25	107.329	-0.07	0.07407
1996	115.13	115.494	-0.36	0.36394
1997	124.16	123.659	0.50	0.49653
1998	130.11	131.824	-1.72	1.71579
1999	138.57	139.989	-1.41	1.41423
2000	146.97	148.154	-1.18	1.18090
2001	153.40	156.319	-2.92	2.91788
2002	162.28	164.484	-2.21	2.20677
2003	168.73	172.649	-3.92	3.91918
2004	183.09	180.814	2.27	2.27388
2005	195.74	188.979	6.76	6.76142

MAE = 2.12

FORECAST ACCURACY MEASURES

Mean Square Error : MSE

Step 1: Calculate Error : Error = Actual - Predicted

Step 2: Square Errors

Step 3: Calculate MSE : MSE = Average of Squared Error

Year	GDP	Predicted	Error	Error Square
1993	94.43	91	3.43	11.73675
1994	100.00	99.165	0.83	0.69722
1995	107.25	107.329	-0.07	0.00549
1996	115.13	115.494	-0.36	0.13245
1997	124.16	123.659	0.50	0.24654
1998	130.11	131.824	-1.72	2.94393
1999	138.57	139.989	-1.41	2.00006
2000	146.97	148.154	-1.18	1.39452
2001	153.40	156.319	-2.92	8.51401
2002	162.28	164.484	-2.21	4.86985
2003	168.73	172.649	-3.92	15.35998
2004	183.09	180.814	2.27	5.17053
2005	195.74	188.979	6.76	45.71683

MSE = 7.60

FORECAST ACCURACY MEASURES

Mean Absolute Percentage Error : MAPE

Step 1: Calculate Error : Error = Actual - Predicted

Step 2: Calculate relative or percentage error : % Error = (absolute(Actual – Predicted) / Actual) x 100 = (absolute Error / Actual) x 100

Step 3: Calculate MAPE

Year	GDP	Predicted	Error	% Error
1993	94.43	91	3.43	3.62813
1994	100.00	99.165	0.83	0.83500
1995	107.25	107.329	0.07	0.06906
1996	115.13	115.494	0.36	0.31611
1997	124.16	123.659	0.50	0.39992
1998	130.11	131.824	1.72	1.31874
1999	138.57	139.989	1.41	1.02056
2000	146.97	148.154	1.18	0.80348
2001	153.40	156.319	2.92	1.90212
2002	162.28	164.484	2.21	1.35988
2003	168.73	172.649	3.92	2.32276
2004	183.09	180.814	2.27	1.24196
2005	195.74	188.979	6.76	3.45428

MAPE = 1.437

FORECAST ACCURACY MEASURES

Mean Absolute Percentage Error : WMAPE

Step 1: Calculate Error : Actual - Predicted

Step 2: Calculate WMAPE :

$$\frac{\sum\left[\left(\frac{|(F-A)|}{A}\right)*100*A\right]}{\sum A}$$

Year	GDP	Predicted	Error
1993	94.43	91	3.43
1994	100	99.165	0.835
1995	107.25	107.329	0.079
1996	115.13	115.494	0.364
1997	124.16	123.659	0.501
1998	130.11	131.824	1.714
1999	138.57	139.989	1.419
2000	146.97	148.154	1.184
2001	153.4	156.319	2.919
2002	162.28	164.484	2.204
2003	168.73	172.649	3.919
2004	183.09	180.814	2.276
2005	195.74	188.979	6.761
Sum	1819.86		27.605

WMAPE = 1.517

FORECAST ACCURACY MEASURES

Exercise: The data on shipments over a periods of time in the chronological order is given below. The forecasts obtained using two different methods are also given below. Identify which forecasting method is more accurate using MAE, MSE , MAPE, WMAPE?

Shipments	Forecast 1	Forecast 2
115	70.333	89.167
132	94.667	112.212
141	115.667	135.258
154	129.333	158.303
171	142.333	181.348
180	155.333	204.394
204	168.333	227.439
228	185	250.485
247	204	273.53
291	226.333	296.576
337	255.333	319.621
391	291.667	342.667

FORECAST PREDICTION INTERVAL

Prediction Interval

Prediction interval : Predicted value $\pm$ z $\sqrt{\text{MSE}}$

where z = width of prediction interval

Prediction Interval	z
90%	1.645
95%	1.960
99%	2.576

FORECAST PREDICTION INTERVAL

Prediction Interval

Example: The data on Yearly average of Indian GDP during 1993 to 2005. The predicted values using a suitable forecasting method is also given.Calculate 95% prediction interval

Year	GDP	Predicted
1993	94.43	91
1994	100.00	99.165
1995	107.25	107.329
1996	115.13	115.494
1997	124.16	123.659
1998	130.11	131.824
1999	138.57	139.989
2000	146.97	148.154
2001	153.40	156.319
2002	162.28	164.484
2003	168.73	172.649
2004	183.09	180.814
2005	195.74	188.979

FORECAST PREDICTION INTERVAL

Prediction Interval

Example: The data on Yearly average of Indian GDP during 1993 to 2005. The predicted values using a suitable forecasting method is also given.Calculate 95% prediction interval

Year	GDP	Predicted	Error	Square Error
1993	94.43	91	3.43	11.73675
1994	100.00	99.165	0.83	0.69722
1995	107.25	107.329	0.07	0.00549
1996	115.13	115.494	0.36	0.13245
1997	124.16	123.659	0.50	0.24654
1998	130.11	131.824	1.72	2.94393
1999	138.57	139.989	1.41	2.00006
2000	146.97	148.154	1.18	1.39452
2001	153.40	156.319	2.92	8.51401
2002	162.28	164.484	2.21	4.86985
2003	168.73	172.649	3.92	15.35998
2004	183.09	180.814	2.27	5.17053
2005	195.74	188.979	6.76	45.71683

MSE	7.60
$\sqrt{MSE}$	2.76
z	1.96
Prediction Interval	5.40

FORECAST PREDICTION INTERVAL

Prediction Interval

Example: The data on Yearly average of Indian GDP during 1993 to 2005. The predicted values using a suitable forecasting method is also given.Calculate 95% prediction interval

Prediction Interval				
Year	GDP	Predicted	Lower Limit	Upper Limit
1993	94.43	91	85.597	96.403
1994	100.00	99.165	93.762	104.568
1995	107.25	107.329	101.926	112.732
1996	115.13	115.494	110.091	120.897
1997	124.16	123.659	118.256	129.062
1998	130.11	131.824	126.421	137.227
1999	138.57	139.989	134.586	145.392
2000	146.97	148.154	142.751	153.557
2001	153.40	156.319	150.916	161.722
2002	162.28	164.484	159.081	169.887
2003	168.73	172.649	167.246	178.052
2004	183.09	180.814	175.411	186.217
2005	195.74	188.979	183.576	194.382

MSE	7.60
√MSE	2.76
z	1.96
Prediction Interval	5.40

FORECAST PREDICTION INTERVAL

Prediction Interval

Example: The data on shipments over a periods of time in the chronological order is given below with the forecasted values. Provide 95% prediction interval?

Shipments	Forecast
115	89.167
132	112.212
141	135.258
154	158.303
171	181.348
180	204.394
204	227.439
228	250.485
247	273.53
291	296.576
337	319.621
391	342.667

R-code

```
model = ses(x)
summary(model)
```

Moving Average Method

Moving Average: The average of successive smaller set of data

Example: The data on shipments over a periods of time in the chronological order is given below. Calculate the forecasts using moving average of length 3?

Period	Shipments
1	123
2	112
3	108
4	118
5	95
6	109
7	122
8	108
9	112
10	116
11	103
12	110

FORECAST METHODS

Moving Average Method

Moving Average: The average of successive smaller set of data

Step1 : Make time series plot

Moving Average Method

Moving Average: The average of successive smaller set of data

Example: The data on shipments over a periods of time in the chronological order is given below. Calculate the forecasts using moving average of length 3?

Period	Shipments		Forecast
1	123		
2	112		
3	108		
4	118	(123+112+108)/3	114.3333
5	95		112.6667
6	109		107
7	122		107.3333
8	108		108.6667
9	112		113
10	116		114
11	103		112
12	110		110.3333
13			109.6667

MAE	5.678
MSE	70.31
MAPE	5.31

FORECAST METHODS

Moving Average Method

Step 1: Take Moving average Length k = 2

Step 2: Calculate moving average of length k

Step 3: Calculate Forecast Accuracy Measures (MAD, MSD or MAPE)

Step 4: Repeat step 2 & 3 with k = 3, 4 - - - , 10 or 12

Step 5: Identify the optimum k . The k with minimum MAD or MSD.

Step 6: Calculate the forecasts as moving average of length optimum k

Step 7: Calculate prediction intervals, if required

Moving Average Method

Exercise 1:The data on yearly income before taxes of a PC manufacturer is given below:. Forecast the income in the coming year using moving average method? Calculate the prediction interval?

Year	Income (Million $)
1997	46.163
1998	46.998
1999	47.816
2000	48.311
2001	48.758
2002	49.164
2003	49.548
2004	48.915
2005	50.315
2006	50.768

Moving Average Method

Exercise 2: The data on monthly sales figures of an electronic component for the last 3 years is given below. Forecast the sales volume for the upcoming month using moving average method?

Month	Sales	Month	Sales	Month	Sales
1	266	13	194	25	339
2	145	14	149	26	440
3	183	15	210	27	315
4	119	16	273	28	439
5	180	17	191	29	401
6	168	18	287	30	437
7	231	19	226	31	575
8	224	20	303	32	407
9	192	21	289	33	682
10	122	22	421	34	475
11	336	23	264	35	581
12	185	24	342	36	646

SINGLE EXPONENTIAL SMOOTHING

Moving Average Method: Issues

Give equal weightage to all the values

Single Exponential Smoothing:

Give more weight to recent values compared to the old values

Single Exponential Smoothing: Methodology

Let $y_1, y_2, - - - y_t$ be the values, then

y_{t+1} estimate = $S_{t+1} = \alpha\, y_t + (1 - \alpha)\, S_t$

where $0 \leq \alpha \geq 1$ and $S_1 = y_1$

SINGLE EXPONENTIAL SMOOTHING

Example: The data on ad revenue from an advertising agency for the last 12 months is given below. Forecast the ad revenue from the agency in the future month using single exponential smoothing method with $\alpha = 0.13$?

Month	Amount	Month	Amount
1	9	7	11
2	8	8	7
3	9	9	13
4	12	10	9
5	9	11	11
6	12	12	10

SINGLE EXPONENTIAL SMOOTHING

Example: Forecasts using single exponential smoothing method with $\alpha = 0.13$?

Month	Amount	Forecasts
1	9	
2	8	9.00
3	9	8.87
4	12	8.89
5	9	9.29
6	12	9.25
7	11	9.61
8	7	9.79
9	13	9.43
10	9	9.89
11	11	9.78
12	10	9.94

Forecast of $y_2 = y_1 = 9.00$

Forecast of $y_3 = \alpha . y_2 + (1 - \alpha)$ (y_2 Forecast) $= 0.13 \times 8 + (1 - 0.13) \times 9 = 8.87$

SINGLE EXPONENTIAL SMOOTHING

Determination of α

Step 1:

Choose $\alpha = 0.1$

Step 2:

Forecast Values

Step 3:

Calculate Errors

Step 4:

Calculate SSE and MSE

Step 5:

Repeat steps 1 to 4 for different values of α

Step 6:

Choose the α with minimum MSE

SINGLE EXPONENTIAL SMOOTHING

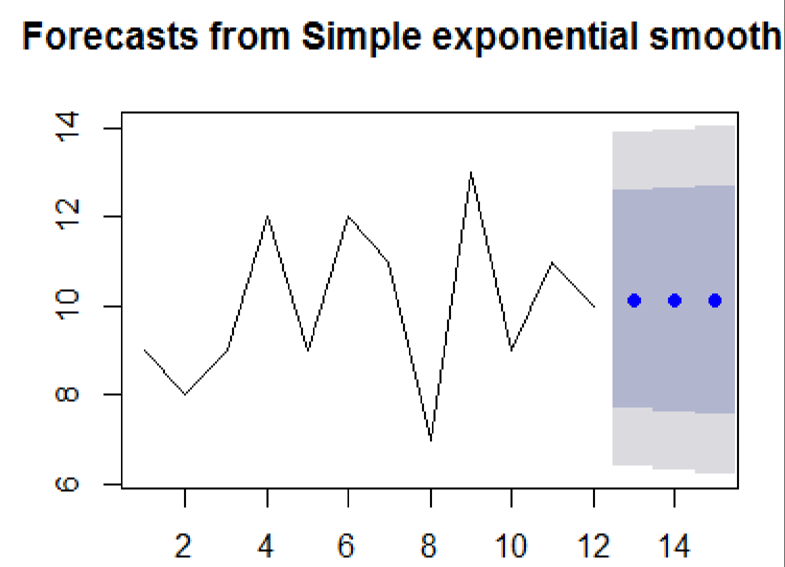
```
amt=ts(amount[,"Amount"])
plot(amt)
fit1=ses(amt,alpha=0.2,initial="simple",h=3)
summary(fit1)
```

```
> amt=ts(amount[,"Amount"])
> plot(amt)
> fit1=ses(amt,alpha=0.2,initial="simple",h=3)
> plot(fit1)
> summary(fit1)

Forecast method: Simple exponential smoothing

Model Information:


Call:
 ses(x = amt, h = 3, initial = "simple", alpha = 0.2)

  Smoothing parameters:
    alpha = 0.2

  Initial states:
    l = 9

  sigma:  1.9125
Error measures:
                    ME     RMSE      MAE      MPE     MAPE
Training set 0.4722112 1.912504 1.465238 1.624006 14.50994
                MASE       ACF1
Training set 0.55578 -0.5428488

Forecasts:
   Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
13       10.13331 7.682334 12.58428 6.384867 13.88175
14       10.13331 7.633795 12.63282 6.310634 13.95598
15       10.13331 7.586181 12.68043 6.237814 14.02880
```

amt	Time-Series [1:12] from 1 to 12: 9 ...
data	Time-Series [1:16] from 2010 to 201...
data1	Time-Series [1:15] from 2010 to 201...
data2	Time-Series [1:16] from 1 to 16: 99...
fit1	List of 9

Files Plots Packages Help Viewer

Zoom Export Clear All



SINGLE EXPONENTIAL SMOOTHING

Example: The data on ad revenue from an advertising agency for the last 12 months is given below. Forecast the ad revenue from the agency in the future month using single exponential smoothing method with best value of α?

Month	Amount	Month	Amount
1	9	7	11
2	8	8	7
3	9	9	13
4	12	10	9
5	9	11	11
6	12	12	10

Holt’s Exponential smoothing

- **Holt’s two parameter exponential smoothing method is an extension of simple exponential smoothing.**
- **It adds a growth factor (or trend factor) to the smoothing equation as a way of adjusting for the trend.**
- **Three equations and two smoothing constants are used in the model.**
  - **The exponentially smoothed series or current level estimate.**

$$L_t = \alpha y_t + (1-\alpha)(L_{t-1} + b_{t-1})$$

  - **The trend estimate.**

$$b_t = \beta(L_t - L_{t-1}) + (1-\beta)b_{t-1}$$

  - **Forecast m periods into the future.**

$$F_{t+m} = L_t + mb_t$$

Holt's Exponential smoothing

- L_t = Estimate of the level of the series at time t
- α = smoothing constant for the data.
- y_t = new observation or actual value of series in period t.
- β = smoothing constant for trend estimate
- b_t = estimate of the slope of the series at time t
- m = periods to be forecast into the future.

- The weight α and β can be selected subjectively or by minimizing a measure of forecast error such as RMSE.
- Large weights result in more rapid changes in the component.
- Small weights result in less rapid changes.

Holt's Exponential smoothing

- **The initialization process for Holt's linear exponential smoothing requires two estimates:**
  - **One to get the first smoothed value for L_1**
  - **The other to get the trend b_1.**
- **One alternative is to set $L_1 = y_1$ and**

$$b_1 = y_2 - y_1$$

$$or$$

$$b_1 = \frac{y_4 - y_1}{3}$$

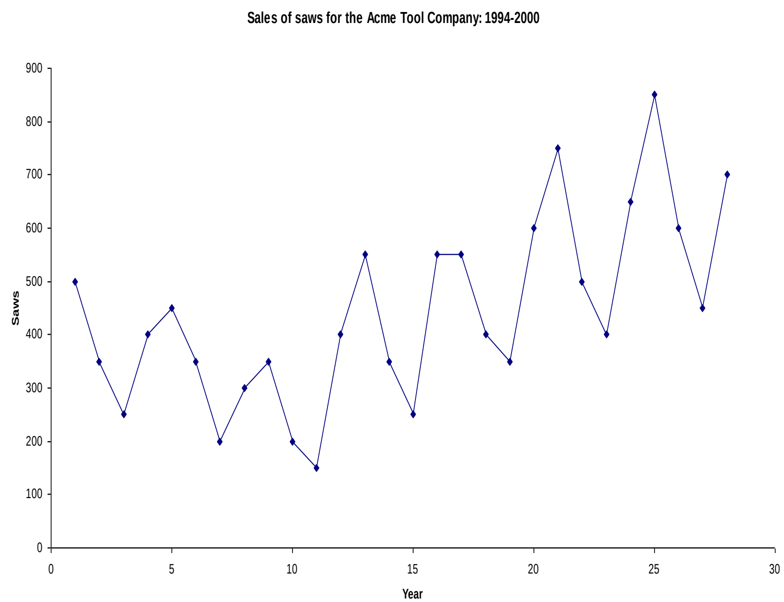
$$or$$

$$b_1 = 0$$

Holt's Exponential smoothing

- **The following table shows the sales of saws for the a tool Company**
- **These are quarterly sales From 1994 through 2000.**

Year	Quarter	t	sales
1994	1	1	500
	2	2	350
	3	3	250
	4	4	400
1995	1	5	450
	2	6	350
	3	7	200
	4	8	300
1996	1	9	350
	2	10	200
	3	11	150
	4	12	400
1997	1	13	550
	2	14	350
	3	15	250
	4	16	550
1998	1	17	550
	2	18	400
	3	19	350
	4	20	600
1999	1	21	750
	2	22	500
	3	23	400
	4	24	650
2000	1	25	850
	2	26	600
	3	27	450
	4	28	700



Examination of the plot shows:

- A non-stationary time series data.
- Seasonal variation seems to exist. Sales for the first and fourth quarter are larger than other quarters.

Holt's Exponential smoothing

1. **The plot of the data shows that there might be trending in the data therefore we will try Holt's model to produce forecasts.**

2. **We need two initial values**
   - **The first smoothed value for L_1**
   - **The initial trend value b_1.**

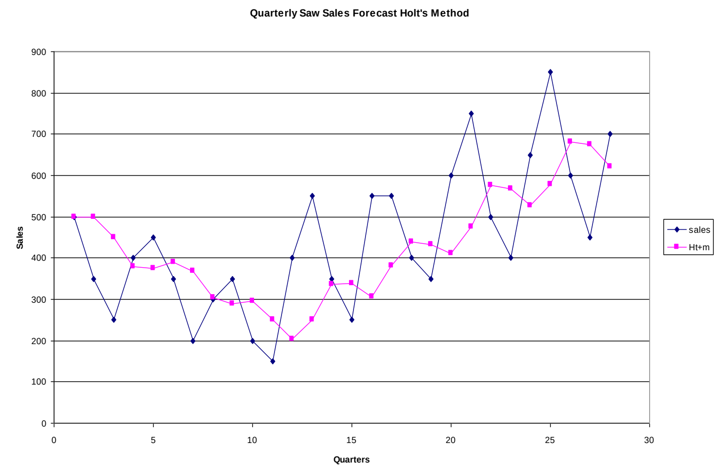
3. **We will use the first observation for the estimate of the smoothed value L_1, and the initial trend value $b_1 = 0$.**

4. **We will use $\alpha = .3$ and $\beta = .1$.**

Holt's Exponential smoothing

Example - Quarterly sales of saws for a tool company

Year	Quarter	t	sales	L_t	b_t	F_{t+m}
1994	1	1	500	500.00	0.00	500.00
	2	2	350	455.00	-4.50	500.00
	3	3	250	390.35	-10.52	450.50
	4	4	400	385.88	-9.91	379.84
1995	1	5	450	398.18	-7.69	375.97
	2	6	350	378.34	-8.90	390.49
	3	7	200	318.61	-13.99	369.44
	4	8	300	303.23	-14.13	304.62
1996	1	9	350	307.38	-12.30	289.11
	2	10	200	266.55	-15.15	295.08
	3	11	150	220.98	-18.19	251.40
	4	12	400	261.95	-12.28	202.79
1997	1	13	550	339.77	-3.27	249.67
	2	14	350	340.55	-2.86	336.50
	3	15	250	311.38	-5.49	337.69
	4	16	550	379.12	1.83	305.89
1998	1	17	550	431.67	6.90	380.95
	2	18	400	427.00	5.74	438.57
	3	19	350	407.92	3.26	432.74
	4	20	600	467.83	8.93	411.18
1999	1	21	750	558.73	17.12	476.75
	2	22	500	553.10	14.85	575.85
	3	23	400	517.56	9.81	567.94
	4	24	650	564.16	13.49	527.37
2000	1	25	850	659.35	21.66	577.65
	2	26	600	656.71	19.23	681.01
	3	27	450	608.16	12.45	675.94
	4	28	700	644.43	14.83	620.61



- RMSE for this application is:
  α = .3 and β= .1
  RMSE = 260.09
- The plot also showed the possibility of seasonal variation that needs to be investigated.

Winter's Exponential smoothing

- Winter's exponential smoothing model is the second extension of the basic Exponential smoothing model
- It is used for data that exhibit both trend and seasonality
- It is a three parameter model that is an extension of Holt's method
- An additional equation adjusts the model for the seasonal component.
- The four equations necessary for Winter's multiplicative method are:
  - The exponentially smoothed series:

$$L_t = \alpha \frac{y_t}{S_{t-s}} + (1-\alpha)(L_{t-1} + b_{t-1})$$

  - The trend estimate:

$$b_t = \beta (L_t - L_{t-1}) + (1-\beta) b_{t-1}$$

  - The seasonality estimate:

$$S_t = \gamma \frac{y_t}{L_t} + (1-\gamma) S_{t-s}$$

Winter's Exponential smoothing

- Forecast *m* period into the future:

$$F_{t+m} = (L_t + mb_t)S_{t+m-s}$$

- L_t = level of series.
- α = smoothing constant for the data.
- y_t = new observation or actual value in period t.
- β = smoothing constant for trend estimate.
- b_t = trend estimate.
- γ = smoothing constant for seasonality estimate.
- S_t =seasonal component estimate.
- m = Number of periods in the forecast lead period.
- s = length of seasonality (number of periods in the season)
- F_{t+m} = forecast for m periods into the future.

Winter's Exponential smoothing

- As with Holt's linear exponential smoothing, the weights α, β, and γ can be selected subjectively or by minimizing a measure of forecast error such as RMSE.
- As with all exponential smoothing methods, we need initial values for the components to start the algorithm.
- To start the algorithm, the initial values for L_t, the trend b_t, and the indices S_t must be set.

Winter’s Exponential smoothing

- To determine initial estimates of the seasonal indices we need to use at least one complete season's data (i.e. s periods).Therefore, we initialize trend and level at period s.

- Initialize level as:

$$L_s = \frac{1}{s}(y_1 + y_2 + \cdots y_s)$$

- Initialize trend as

$$b_s = \frac{1}{s}(\frac{y_{s+1} - y_1}{s} + \frac{y_{s+2} - y_2}{s} + \cdots + \frac{y_{s+s} - y_s}{s})$$

- Initialize seasonal indices as:

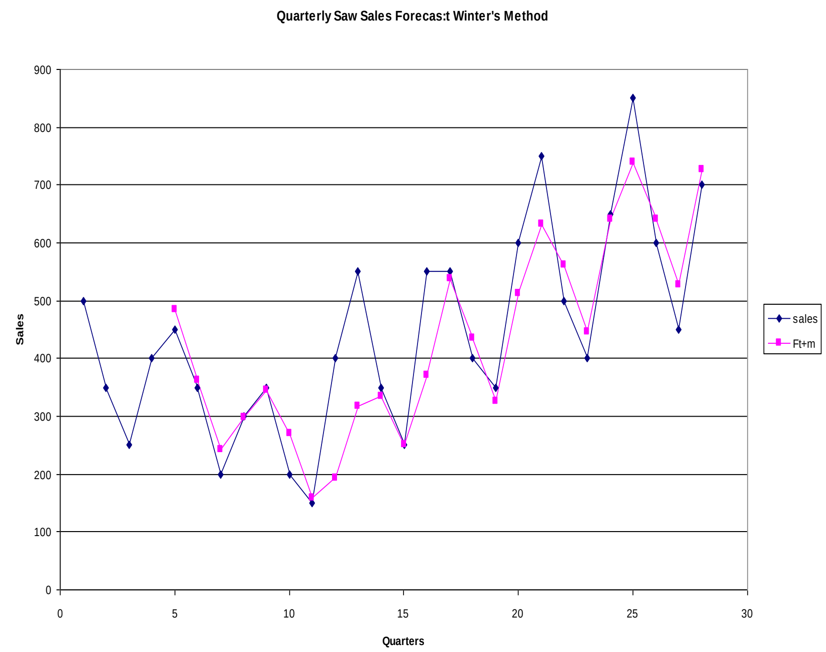
$$S_1 = \frac{y_1}{L_s}, S_2 = \frac{y_2}{L_s}, \cdots, S_s = \frac{y_s}{L_s}$$

Winter's Exponential smoothing

- We will apply Winter's method to Tool company sales. The value for α is .4, the value for β is .1, and the value for γ is .3.
- The smoothing constant α smoothes the data to eliminate randomness.
- The smoothing constant β smoothes the trend in the data set.
- The smoothing constant γ smoothes the seasonality in the data.
- The initial values for the smoothed series L_t, the trend b_t, and the seasonal index S_t must be set.

Winter's Exponential smoothing

Year	Quarter	t	sales	L_t	b_t	S_t	F_{t+m}
1994	1	1	500			1.333	
	2	2	350			0.933	
	3	3	250			0.667	
	4	4	400	375	-12.5	1.067	
1995	1	5	450	396.9667	-9.05333	1.273	483.3333
	2	6	350	372.3747	-10.6072	0.935	362.0524
	3	7	200	296.7938	-17.1046	0.669	241.1783
	4	8	300	287.3869	-16.3348	1.060	298.3352
1996	1	9	350	302.1219	-13.2278	1.239	345.161
	2	10	200	252.9623	-16.821	0.892	270.2048
	3	11	150	201.4173	-20.2934	0.692	157.9377
	4	12	400	268.2504	-11.5807	1.189	191.9611
1997	1	13	550	373.5062	0.102908	1.309	317.9958
	2	14	350	363.8087	-0.87713	0.913	333.2237
	3	15	250	317.4823	-5.42206	0.720	251.002
	4	16	550	406.7605	4.047961	1.238	371.1103
1998	1	17	550	465.9614	9.563264	1.270	537.7528
	2	18	400	444.9496	6.505758	0.909	434.1286
	3	19	350	410.5851	2.418728	0.760	325.2062
	4	20	600	487.3071	9.84905	1.236	511.3412
1999	1	21	750	597.7855	19.91199	1.266	631.5942
	2	22	500	570.255	15.16774	0.899	561.3363
	3	23	400	510.9496	7.720431	0.766841	444.9085
	4	24	650	570.7076	12.92419	1.206915	641.1016
2000	1	25	850	689.6728	23.52829	1.255716	738.6906
	2	26	600	667.561	18.96428	0.899057	641.2886
	3	27	450	591.6084	9.472591	0.764981	526.4561
	4	28	700	640.1658	13.38107	1.172881	725.4539



- fit1=hw(a10,seasonal="additive")
- fit2=hw(a10,seasonal="multiplicative")
- summary(fit1)
- Plot(fit1)

- RMSE for this application is:

  $\alpha = 0.4$, $\beta = 0.1$, $\gamma = 0.3$ and RMSE = 83.36

- Note the decrease in RMSE.

FORECAST METHODS

Time Series Decomposition

Step 1: Draw Time Series Plot of the data

Step 2: If the plot shows a trend as well as cyclic pattern

Step 3: Estimate the forecast values using trend line equation

Decomposition Models

Forecast = F(Seasonal Effect, Trend, Error)

Additive Decomposition

Forecast = Seasonal Effect + Trend Effect + Error

Multiplicative Decomposition

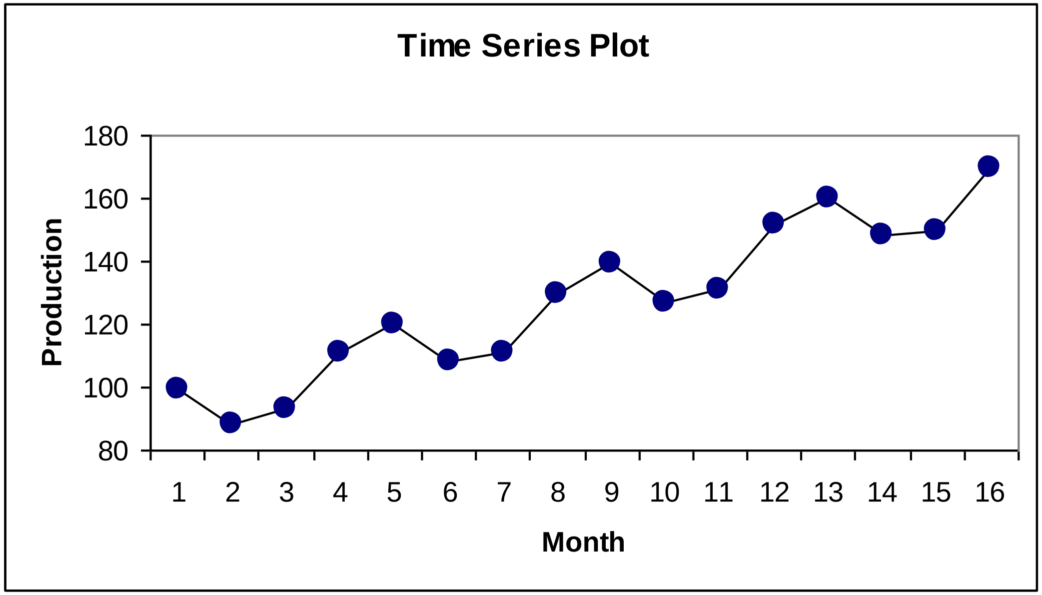
Forecast = Seasonal Effect x Trend Effect x Error

Time Series Decomposition

Time Series Decomposition: Additive

Example: The quarterly manganese production data is given below. Fit a time series model additive decomposition?

Period	Quarter	Production
1	1	99
2	2	88
3	3	93
4	4	111
5	1	120
6	2	108
7	3	111
8	4	130
9	1	139
10	2	127
11	3	131
12	4	152
13	1	160
14	2	148
15	3	150
16	4	170



Remark: There is a trend & seasonality (quarterly) pattern

Time Series Decomposition

Time Series Decomposition: Additive

Example: The quarterly manganese production data is given below. Fit a time series model using additive decomposition?

The Model

Forecast = 85.1938 +4.95515*time

Seasonal Indices

Quarter	Seasonal Index
1	9.78125
2	-6.84375
3	-8.59375
4	5.65625

Time Series Decomposition

Time Series Decomposition: Additive

Example: The quarterly manganese production data is given below. Fit a time series model using additive decomposition?

Period	Quarter	Production	Prediction		Seasonal Index	Seasonal Adjusted Prediction
1	1	99	85.1932+4.95515x1	90.14835	9.78125	90.14835+9.78125 = 99.9296
2	2	88	85.1932+4.95515x2	95.1035	-6.84375	95.1035 - 6.84375 = 88.25915
3	3	93		100.0587	-8.59375	91.4649
4	4	111		105.0138	5.65625	110.67005
5	1	120	85.1932+4.95515x5	109.969	9.78125	109.969+9.78125 = 119.7502
6	2	108		114.9241	-6.84375	108.08035
7	3	111		119.8793	-8.59375	111.2855
8	4	130		124.8344	5.65625	130.49065
9	1	139		129.7896	9.78125	139.5708
10	2	127		134.7447	-6.84375	127.90095
11	3	131		139.6999	-8.59375	131.1061
12	4	152		144.655	5.65625	150.31125
13	1	160		149.6102	9.78125	159.3914
14	2	148		154.5653	-6.84375	147.72155
15	3	150		159.5205	-8.59375	150.9267
16	4	170		164.4756	5.65625	170.13185

Time Series Decomposition

Time Series Decomposition: Multiplicative

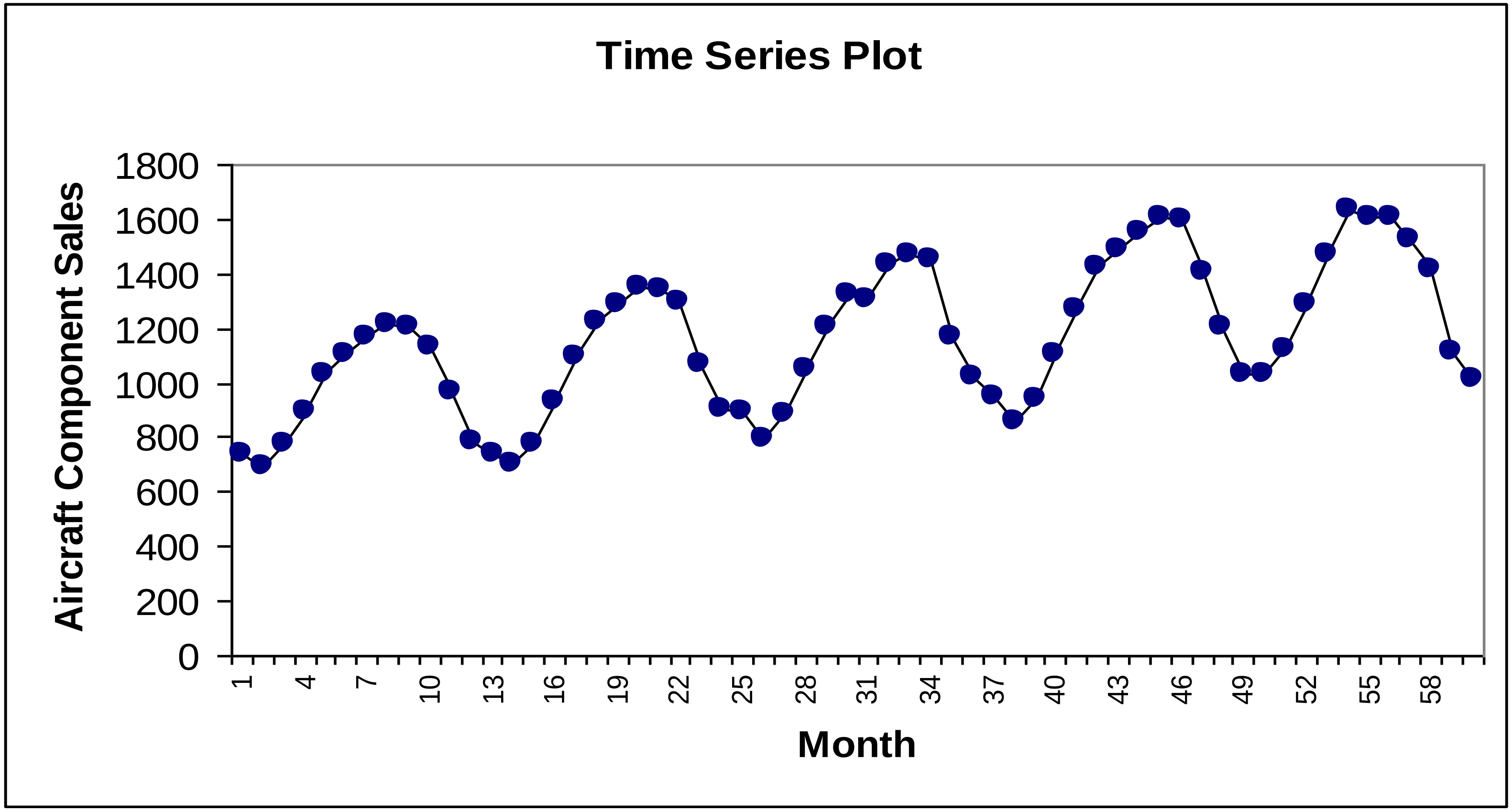
Example: The data on monthly jacket sales is given below. Fit a forecasting model using Multiplicative decomposition?

Month	Sales	Month	Sales	Month	Sales	Month	Sales	Month	Sales
1	742	13	741	25	896	37	951	49	1030
2	697	14	700	26	793	38	861	50	1032
3	776	15	774	27	885	39	938	51	1126
4	898	16	932	28	1055	40	1109	52	1285
5	1030	17	1099	29	1204	41	1274	53	1468
6	1107	18	1223	30	1326	42	1422	54	1637
7	1165	19	1290	31	1303	43	1486	55	1611
8	1216	20	1349	32	1436	44	1555	56	1608
9	1208	21	1341	33	1473	45	1604	57	1528
10	1131	22	1296	34	1453	46	1600	58	1420
11	971	23	1066	35	1170	47	1403	59	1119
12	783	24	901	36	1023	48	1209	60	1013

Time Series Decomposition

Time Series Decomposition: Multiplicative

Example: The data on monthly jacket sales is given below. Fit a forecasting



Time Series Decomposition

Time Series Decomposition: Multiplicative

Example: The data on monthly jacket sales is given below. Fit a forecasting model using Multiplicative decomposition?

The Model: Forecast = 931.374 + 7.56513*t

Month	Seasonality Index
1	0.76732
2	0.70541
3	0.77146
4	0.91119
5	1.0465
6	1.14901
7	1.17224
8	1.23201
9	1.23527
10	1.1934
11	0.98471
12	0.83149

Time Series Decomposition

Time Series Decomposition: Multiplicative

Period	Month	Sales	Prediction	Period	Month	Sales	Prediction
1	1	742	720.47	31	7	1303	1366.71
2	2	697	667.67	32	8	1436	1445.71
3	3	776	736.02	33	9	1473	1458.88
4	4	898	876.23	34	10	1453	1418.47
5	5	1030	1014.27	35	11	1170	1177.86
6	6	1107	1122.31	36	12	1023	1000.88
7	7	1165	1153.87	37	1	951	929.45
8	8	1216	1222.02	38	2	861	859.79
9	9	1208	1234.6	39	3	938	946.13
10	10	1131	1201.79	40	4	1109	1124.39
11	11	971	999.07	41	5	1274	1299.27
12	12	783	849.91	42	6	1422	1435.24
13	1	741	790.13	43	7	1486	1473.12
14	2	700	731.71	44	8	1555	1557.55
15	3	774	806.06	45	9	1604	1571.02
16	4	932	958.95	46	10	1600	1526.81
17	5	1099	1109.27	47	11	1403	1267.25
18	6	1223	1226.62	48	12	1209	1076.36
19	7	1290	1260.29	49	1	1030	999.11
20	8	1349	1333.87	50	2	1032	923.82
21	9	1341	1346.74	51	3	1126	1016.16
22	10	1296	1310.13	52	4	1285	1207.11
23	11	1066	1088.47	53	5	1468	1394.28
24	12	901	925.39	54	6	1637	1539.55
25	1	896	859.79	55	7	1611	1579.54
26	2	793	795.75	56	8	1608	1669.4
27	3	885	876.09	57	9	1528	1683.16
28	4	1055	1041.67	58	10	1420	1635.15
29	5	1204	1204.27	59	11	1119	1356.65
30	6	1326	1330.93	60	12	1013	1151.84

Remark:

In Multiplicative model, the seasonal adjustment is done by multiplying the corresponding seasonality index

```
➢ fit=decompose(try,type="multiplicative")
➢ fit=decompose(try,type="additive")
> summary(fit)
> plot(fit)
> print(fit)
```

Time Series Decomposition

Time Series Decomposition: Multiplicative

Exercise: The sales data on quarterly exports is given for 6 years. Fit a suitable forecasting model

Year	Quarter	Period	Exports	Year	Quarter	Period	Exports
1	1	1	362	4	1	13	544
	2	2	385		2	14	582
	3	3	432		3	15	681
	4	4	341		4	16	557
2	1	5	382	5	1	17	628
	2	6	409		2	18	707
	3	7	498		3	19	773
	4	8	387		4	20	592
3	1	9	473	6	1	21	627
	2	10	513		2	22	725
	3	11	582		3	23	854
	4	12	474		4	24	661

FORECAST METHODS

Auto Regressive Integrated Moving Average Models (ARIMA (p,d,q))

Widely used and very effective modeling approach

Proposed by George Box and Gwilym Jenkins

Also known as Box – Jenkins model or ARIMA(p,d,q)

where

p: number of auto regressive (AR) terms

q: number of moving average (MA) terms

d: level of differencing

FORECAST METHODS

Auto Regressive Integrated Moving Average Models (ARIMA (p,d,q))

General Form

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + - - - + \theta_1 e_{t-1} + + \theta_2 e_{t-2} - - - -$$

Where

c: constant

$\phi_{1,}\ \phi_{2,}\ \theta_{1,}\ \theta_2$, - - - are model parameters

$e_{t-1} = y_{t-1} - s_{t-1}$, e_t are called errors or residuals

s_{t-1} : predicted value for the t-1th observation (y_{t-1})

FORECAST METHODS

Auto Regressive Integrated Moving Average Models (ARIMA (p,d,q))

Step 1:

Draw time series plot and check for trend, seasonality, etc

Step 2:

Draw Auto Correlation Function (ACF) and Partially Auto Correlation Function (PACF) graphs to identify auto correlation structure of the series

Step 3:

Check whether the series is stationary using unit root test (ADF test, KPSS test)

If series is non stationary do differencing or transform the series

FORECAST METHODS

Auto Regressive Integrated Moving Average Models (ARIMA (p,d,q))

Step 4:

Identify the model

Use Hannan-Rissanen procedure to automatically identify the best values of p,d,q, or the AR and MA terms in the model.

The best model is the one which minimizes Akaike Info Criterion (AIC)

Step 5:

Estimate the model parameters using maximum likelihood method (MLE)

Auto Regressive Integrated Moving Average Models (ARIMA (p,d,q))

Step 6:

Do model diagnostic checks

The errors or residuals should be white noise and should not be auto correlated

Do Portmanteau and Ljung & Box tests. If p value > 0.05, then there is no autocorrelation in residuals and residuals are purely white noise.

The model is a good fit

FORECAST METHODS

Auto Regressive Integrated Moving Average Models (ARIMA (p,d,q))

Example: The data daily revenues is given below. Fit Forecasting model?

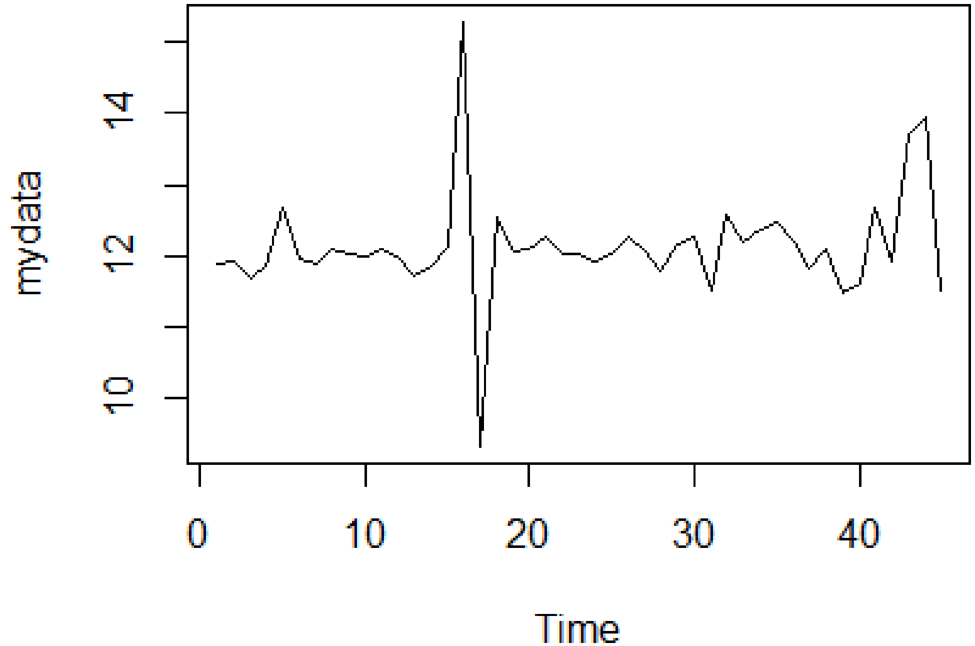
SL No	Data	SL No	Data	SL No	Data
1	11.9	16	15.28	31	11.51
2	11.94	17	9.33	32	12.56
3	11.69	18	12.54	33	12.2
4	11.86	19	12.07	34	12.38
5	12.69	20	12.08	35	12.46
6	11.95	21	12.26	36	12.21
7	11.9	22	12.03	37	11.83
8	12.08	23	12.04	38	12.08
9	12.03	24	11.93	39	11.48
10	11.99	25	12.02	40	11.63
11	12.11	26	12.27	41	12.68
12	11.98	27	12.07	42	11.93
13	11.71	28	11.77	43	13.7
14	11.87	29	12.16	44	13.95
15	12.12	30	12.26	45	11.5

```
mydata=ts(ARIMA1[,"Rev."],frequency=1)
```

FORECAST METHODS

Auto Regressive Integrated Moving Average Models (ARIMA (p,d,q))

Step 1: Time Series Plot plot(mydata)



Auto Regressive Integrated Moving Average Models (ARIMA (p,d,q))

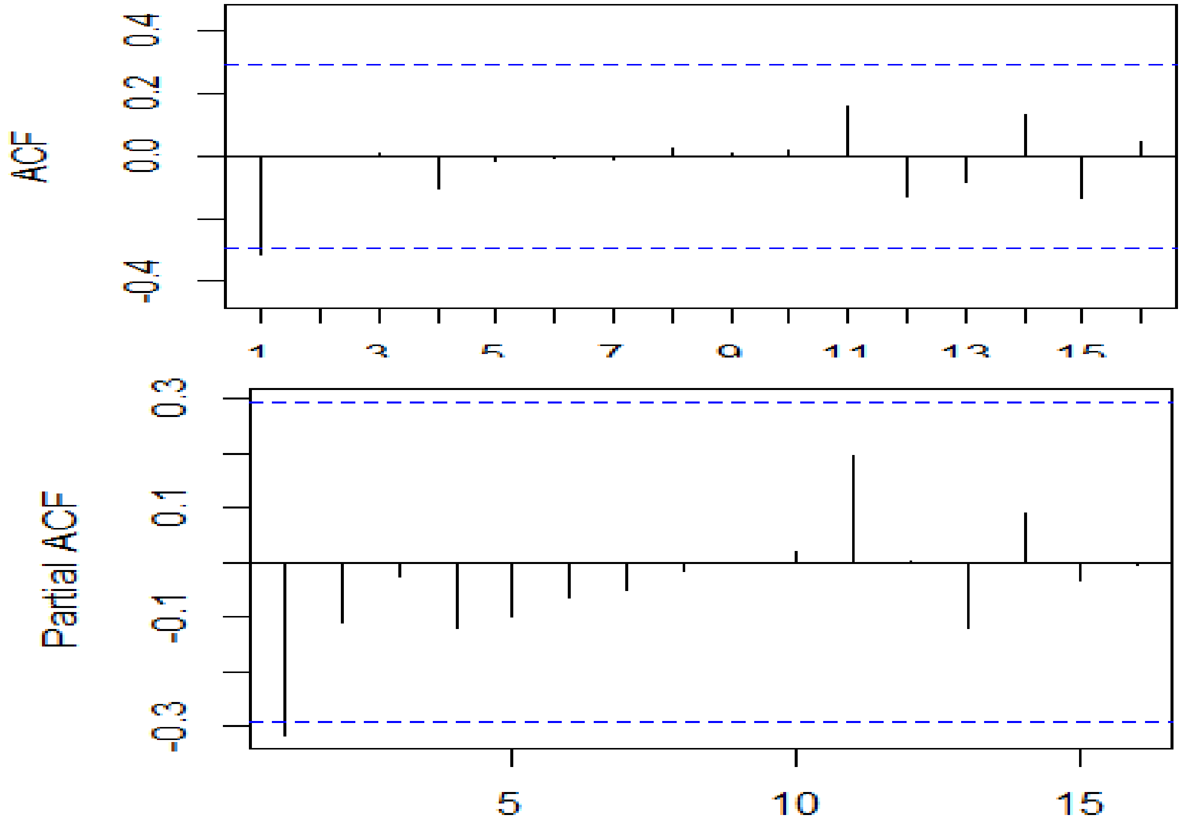
Step 2: Descriptive Statistics

Statistic	Value
Mean	12.134
SD	0.7786
Minimum	9.33
Maximum	15.8

FORECAST METHODS

Auto Regressive Integrated Moving Average Models (ARIMA (p,d,q))

Step 3: Draw ACF & PACF Graphs



Acf(mydata)
Pacf(mydata)

Remark: Only ACF and PACF at lag 1 is significantly higher than 95% confidence limits. Series appears to be stationary

Auto Regressive Integrated Moving Average Models (ARIMA (p,d,q))

Step 4: Do ADF/KPSS test to check the whether the series is stationary

Statistic	Value
ADF Statistic	-3.6273
p-Value	0.04

Statistic	Value
KPSS Statistic	0.1642
p-Value	0.10

Remark: Since ADF statistic < 5% critical value, the series is stationary

```
adf.test(mydata,alternative="stationary")
kpss.test(mydata)
```

FORECAST METHODS

Auto Regressive Integrated Moving Average Models (ARIMA (p,d,q))

Step 5: Identification of parameters

Criteria	Model
Akaike Info Criterion (AIC):	p=0, q=1
Hannan-Quinn Criterion:	p=0, q=1
Schwarz Criterion:	p=0, q=1

Conclusion: All the 3 criteria suggests that the model is p=0, q=1 or MA(1)

```
auto.arima(mydata)
```

Auto Regressive Integrated Moving Average Models (ARIMA (p,d,q))

Step 5: Identification of parameters

Model		Log likelihood	AIC
p=1,q=0	AR(1)	-50.252152	104.504
p=0,q=1	MA(1)	-49.896639	103.793
p=1,q=1	ARMA(1,1)	-49.060318	104.121

Conclusion: The best model which minimizes AIC is p=0, q=1 or MA(1)

Auto Regressive Integrated Moving Average Models (ARIMA (p,d,q))

Step 6: Estimation of parameters

	Coefficients	Std. Errors
MA1	- 0.377	0.1651
Constant	12.1349	0.0689

The model is $y_t = a + \theta_1 e_{t-1}$

$$y_t = 12.135 - 0.3778 e_{t-1}$$

```
model=arima(mydata,order=c(0,0,1))
Summary(model)
Forecast(model,h=3)
```

Auto Regressive Integrated Moving Average Models (ARIMA (p,d,q))

Step 7: Model diagnostics

Portmanteau and Ljung & Box Tests

```
res=residuals(model)
Acf(res)
Box.test(res,lag=10,fitdf=0,type="Lj")
Portest(res)
```

Statistic	Value	p value
Portmanteau	0.6409	0.9381
Ljung & Box	1.8247	0.9975

Since the p values for both test > 0.05, The model fits the data

The residuals are not auto correlated

The residuals are white noise

Auto Regressive Integrated Moving Average Models (ARIMA (p,d,q))

Exercise 1: The number of visitors to a web page given below. Develop a model to predict the daily number of visitors?

SL No.	Data	SL No.	Data
1	259	16	416
2	310	17	248
3	268	18	314
4	379	19	351
5	275	20	417
6	102	21	276
7	139	22	164
8	60	23	120
9	93	24	379
10	45	25	277
11	101	26	208
12	161	27	361
13	288	28	289
14	372	29	138
15	291	30	206

Auto Regressive Integrated Moving Average Models (ARIMA (p,d,q))

Exercise 2: The following table gives the data on sales of a electro magnetic component. Develop a forecasting methodology?

Period	Data	Period	Data
1	4737	16	4405
2	5117	17	4595
3	5091	18	5045
4	3468	19	5700
5	4320	20	5716
6	3825	21	5138
7	3673	22	5010
8	3694	23	5353
9	3708	24	6074
10	3333	25	5031
11	3367	26	5648
12	3614	27	5506
13	3362	28	4230
14	3655	29	4827
15	3963	30	3885