

Training Program on
Statistical Techniques
for
Data Mining & Business Analytics

10-12 February and 3-5 March 2017

Sl No.	Topics	Page No.	Sl No.	Topics	Page No.
1	Introduction to Business Analytics	3	11	Predictive Analytics	209
2	Preliminary Analysis of Data	19	11.1	Correlation & Regression	211
3	Introduction to R and R-studio	35	11.2	Modelling Non linear relations	271
4	Descriptive Statistics using R	39	11.3	Binary Logistic Regression	285
5	Data Preprocessing	63	11.4	Tree based Models	297
6	Fundamentals of Probability and Probability Distribution	75	12	Multi Response Scoring Methods	331
7	Test of Hypothesis	115	13	Market Basket Analysis	351
8	Normality Test	157	14	Factor Analysis	375
9	Analysis of Variance	163	15	Cluster Analysis	393
10	Cross Tabulation and Chi-square test	183	16	Naive Bayes Classifier	405

Introduction to Business Analytics

Introduction to Business Analytics

- **An overview of Business Analytics – the types of problems solved**
- **The constituents and the process of implementing BA solutions**
- **Classification of Analytics problems**
- **The way forward – the organization of the course**

Some Examples

1. An automobile manufacturer wants to understand how the fault and failure related data captured through the sensors may be used to classify the condition of vehicles so that preventive maintenance may be carried out optimally. Similar situations are applicable to large manufacturers having many machines, e.g. miners, aircraft manufacturers.
2. Insurers may wish to classify drivers as very risky, risky, safe etc. on the basis of their driving habits so that insurance premium may be fixed intelligently
3. A company engaged in oil exploration may need to estimate the time and expenses of drilling under different geological conditions before taking up a drilling assignment
4. A company in any segment may wish to forecast the total demand based on past demands as well as past and current economic conditions
5. Manufacturers of consumer electronics may need to understand the sentiment of people communicating over social media about their products
6. A large retailer may like to understand the impact of a natural disaster like a hurricane on purchase behaviour
7. An e-commerce company may want to know the impact of making changes in the portal or sales policy on the quantum of sales
8. Credit card as well as health insurance companies may wish to identify fraudulent transactions so that appropriate actions may be initiated
9. A retailer may like to suggest additional products a customer may be willing to buy on the basis of the current as well as past surfing data

Notes

- Business Analytics – variously called as data science or statistical learning – is a set of methods to arrive at quantitative solutions to problems of business interest
- This subject has assumed tremendous importance in the recent past as the data availability is constantly on the rise and there is widespread belief that the existing data may be fruitfully analyzed to arrive at hitherto unknown insights

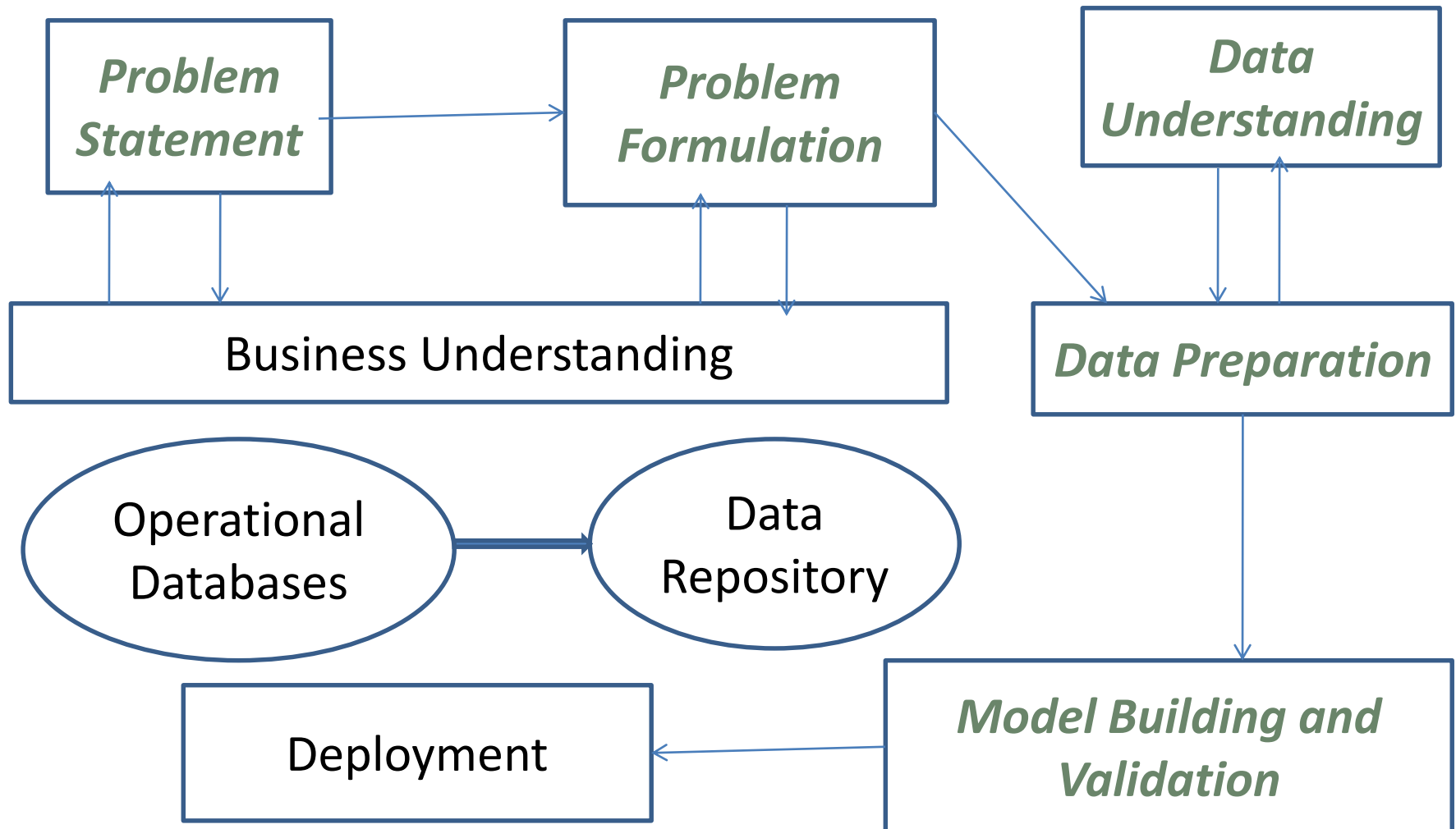
Definition of Business Analytics (BA)

1. BA is the science and the art of improving business functions using data and analytical techniques
2. It is a science since it uses theories of probability, statistics, data mining techniques and a well defined process
3. It is an art because, like a brilliant painter, the analyst has to draw from a diverse pallets of colours (data sources) to find the perfect mix that will yield actionable results
4. It is also an art as the analyst must have a deep level of creativity and business understanding to be able to clearly identify the problem, understand the implementation challenges and effectively communicate the proposed solution. As the saying goes – in business analytics problems will often have to be taken rather than being given

Note: The solution of analytics problems will often come in the following forms

- Insights that may be acted upon (or stop unnecessary actions)
- Models or solutions that may be used to improve effectiveness of business functions
- Automatic solutions embedded in software systems

Business Analytics Process



Types of Analytics Problems

- Analytics problems may be classified from two perspectives, namely
 - Method of analyses
 - Type of business problem

Two Broad Types from Methods Perspective

- Supervised learning
 - Understanding the behaviour of a target (response / dependent / Y) variable as a set of inputs (independent / explanatory / X) vary
 - Typically attempts are made to develop a function to estimate the target
 - These methods are often called dependency analyses
- Unsupervised learning
 - Discovering associations and patterns among a set of input measures. After patterns are found, the analyst is responsible for finding how to interpret and use them.
 - These analyses do not attempt to estimate some Y on the basis of X variables. Rather, attempts are made to understand relationships / patterns of X variables.
 - These methods are often called inter-dependency analyses

Examples of Supervised Learning

- Predict whether a patient, hospitalized due to a heart attack, will have a second heart attack. The prediction is to be based on demographic, diet and clinical measurements for that patient.
- Predict the price of a stock in 6 months from now, on the basis of company performance measures and economic data.
- Predict whether a particular credit card transaction is fraudulent. The prediction is to be based on past transaction history, transaction type, reputation of the merchants involved and other similar variables
- Identify the impact of different variables like price, relative brand position, general economic condition, level of competition, and product type (luxury / necessity...) on the demand of a particular product during a given period

Examples of Unsupervised Analytics

- Find typical profile of employees who quit quickly
- Find products that are usually sold together
- Group cities with respect to their characteristics
- Develop a scale to measure brand position

Analytic Tasks from Business Perspective

1. Hypotheses testing
2. Classification and class probability estimation
3. Value estimation, explanatory and causal models
4. Discovering dimensions, and construction and validation of measures
5. Profiling – understanding behavioural pattern of individual entities
6. Associations and co-occurrence grouping
7. Exploration of phenomenon and understanding trends
8. Link prediction
9. Constrained optimizations (primarily LP, its variants and network optimization)

Most business problems can be solved using a combination of these tasks. As an analyst you should be in a position to break a problem in terms of these tasks.

In this session we will look at only the first three points.

Hypotheses Testing

- Hypotheses are statements about a given phenomenon, e.g. increasing number of years of education increases earning potential; design A produces a lower defect rate compared to design B; a particular design of a web page leads to more conversion compared to another
- Hypotheses testing consists of determining the plausibility of the statements on the basis of data

Classification and Class Probability Estimation

- There are situations where the target is classified, e.g. whether a particular credit card transaction is fraudulent or not; whether a customer will renew her contract or not; whether a sales bid will be won, lost or abandoned by the customer; how to classify a loan application as low, high or medium risk
- The problem is to allocate the target variable to one of the classes based on the value of some explanatory variable(s)
- In most cases the probability that the target will belong to different classes is first estimated. An allocation to a particular class is made on the basis of the estimated probabilities

Value Estimation

- Some business problems require estimating the value of a target variable rather than classifying the same.
- Some examples of value estimation are – finding the lifetime value of a customer; estimating the effort required to complete a software development project; finding the total number of cheques that may arrive for processing...
- The value needs to be estimated based on certain explanatory variables and hence this task comes under the broad class of supervised analytics (dependency analyses)

Relationship Between Fundamental Tasks and Techniques

Sl. No.	Fundamental Task	Statistical / Data Mining Techniques
1	Phenomenon Understanding	Descriptive Statistics , EDA, hypothesis testing, graphical analysis and data visualization, contingency tables
2	Classification	Logistic Regression, Discriminant Analysis, Decision Trees – CART and CHAID, ANN, Support Vector Machine
3	Value Estimation	Table Lookup, Naive Bayesian, Nearest Neighbor, Regression models – MLR and its variants including shrinkage methods, Count Regression and zero inflated models, Cox Regression, Survival Analysis, different non-parametric methods, ANN

Summary

- In the current scenario, ability to extract meaningful insights from the data is of utmost importance. Many organizations have gained significantly by using analytics and non-usage may be costly from a strategic perspective
- Business Analytics consists of three major areas. In this course we are covering the data science aspect in detail. Business analytics problems may be looked at from two perspectives – methods and business problems
- Statistical learning problems may be divided into two broad classes from methods perspective – supervised and unsupervised
- Supervised learning consists of building models to establish relationships between one or more dependent variables (target) Y and a set of input (explanatory) variables. Unsupervised learning finds patterns of association of the inputs
- Business problems may be divided into nine major classes and real life problems may be expressed as their combination
- An expert in Business Analytics must know the different techniques of supervised and unsupervised learning. At the same time s/he should be in a position to construct a business problems in terms of the fundamental tasks.

Preliminary Analyses of Data

Basics of Measurement

- Measurement is an activity that allocates a value or a symbol to a physical attribute of an entity
- Notes:
 1. Ensure that the physical attribute is clearly understood and the measurement truly pertains to this attribute. Failure to do so leads to problem of *representational validity*. Be careful of derived and subjective measures.
 2. Make sure that the entity being measured is clearly identified
 3. Make sure that the measurement method is defined, understood and followed. This is known as *operational definition*. Failure to define measures operationally may lead to serious problems.

Valid and Invalid Measures

- Measurements may be
 - Invalid from a representational perspective
 - Not defined operationally
 - A construct (unobserved like skill / knowledge / satisfaction...) that is not being measured correctly (construct validity – has a similarity with representational validity)
 - A construct that is not being covered from all aspects, e.g. measurement of employee satisfaction may not cover all aspects (content validity)
 - Unreliable because of the method / instrument

Data Quality

- Some of the data quality parameters are representational, construct and content validity
- Data quality may suffer from lack of operational definition as well as issues related to its reliability
- A few other characteristics are
 - Timeliness (whether data are made available on time)
 - Completeness (whether all aspects / fields of the data are available)
 - Accuracy (whether the data reflects the situation accurately)
 - Note: Problem of accuracy may arise because of operational definition, internal political reasons as well as lack of motivation
 - Relevance (whether the data may be used for analyses / decision making)

Measurement Scales

- Measurement is normally carried out in four different scales
 - **Nominal:** Only names (labels) with no implicit ordering, e.g. the department you belong to, the state you come from, the city where a customer is located
 - **Ordinal:** Categorical measures with an implicit ordering, e.g. level of satisfaction measured in a 5 point scale, socio-economic status
 - **Interval:** This is a numeric scale with arbitrary zero. Examples of measurements carried out in this scale are temperature measured in Fahrenheit or Centigrade, calendar year, measurements carried out in comparative scales...
 - **Ratio scale:** This is the highest scale (i.e. contains maximum information). This is a numeric scale with non-arbitrary zero. All arithmetic operations are permissible and ratios make sense. Examples – length, time, number of defects, value of purchase...

Exercise

- Classify the following variables in the appropriate measurement scale
 - Gender of a person
 - Number of accidents in Mumbai on a given day
 - Level of intelligence measured using some IQ test (questionnaire based test)
 - Unemployment rate measured as the proportion of unemployed persons in the set of eligible workers (who should be in the job market)
 - Complexity of a computer program measured in three point scale (high, medium, low)
 - Document scrutinized during an hour
 - Viscosity of a liquid

- An analyst needs to make a preliminary assessment of the data before taking up the job of analyses. Typical questions are
 - Are the data collected manually / automatically?
 - Who collects the data? What is his / her motivation?
 - Is the data collection methodology clearly defined?
 - What characteristic is being measured? Is it truly represented?
 - What is the ultimate objective? How will it be met?
 - Can the objective change or evolve? Would it be possible to address the issues in case of a change?

Understanding Interval and Ratio Scale Data

- To start with preliminary analyses are carried out to understand the data (phenomenon / variables)
 - Find the central tendency – usually the mean and the median. Also find the percentiles to get an idea about how the data are distributed over its range
 - Assess the variation. There are a number of measures like the standard deviation, range and inter-quartile range
 - Understand the distributional pattern. This pattern tells us how the data are distributed over its range. We use frequency distribution and histograms to understand this pattern

Note

Measures of central tendency and dispersion are often used to compare random variables. Suppose we are looking at the productivity of software development projects. Suppose a tool vendor claims that usage of their tool will increase software development productivity. In order to test this claim we may have to look at mean or median productivity with and without the tool. This approach is applicable for categorical data as well.

A Note on Measures of Central Tendency

- While average (mean) is easy to calculate and generally produces stable results, it may be impacted by the presence of extreme values. In BA extreme values may be much larger than the usual values and hence it is important to be careful about the presence of such values
 - Note: Taleb has stated that in business scenario we often have *scalable* variables. While physical variables are limited,, the scalable variables like sale, income etc. may be very, very large
- Trimmed Mean: Calculated by removing a percentage of data from both ends of the data set. A trimmed mean is, therefore, the arithmetic average after x-percentage of values have been removed from the highest and lowest ends of the data set
- Winsorized mean is computed by replacing the x-percentage highest and lowest values by the next adjacent value

List of Measures

- Central Tendency (Location):
 - Mean / average
 - Median
 - Mode (most frequently occurring value)
- Dispersion (variation)
 - Standard deviation
 - Range
 - Inter quartile range
- Measures of position
 - Typically measured through percentiles
 - The p^{th} percentile: Values where at least $p\%$ of items are \leq the value and $(100 - p)\%$ are greater (need not be unique)
 - Some typical measures are quartiles (4 portions), quintiles (5 portions), deciles (10 portions)
- Measures of shape
 - Skewness: the degree to which the distribution of data for a variable is largely to one side of the mean
 - Kurtosis: The degree to which the distribution of the data is closely arranged around the mean

Example

Suppose an IT service company is studying its profitability at project level by

- Revenue productivity (**RP**) (revenue earned per man month of effort) and
- Project margin (**PM**) (computed as the ratio of gross profit and gross revenue).

The operations team state that the average RP and PM are US\$ 9325 and 62.10% respectively. You are aware that the corresponding figures in the industry are US\$ 9100 and 55% respectively. Does it give you adequate comfort?

Average alone may not give a good idea about the situation. The following table would give the profile of the variables being considered

Variable	25 th Percentile	50 th Percentile	Average	75 th Percentile
RP (US\$)	4945	6079	9325	7502
PM (%)	55.45	62.56	62.10	70.34

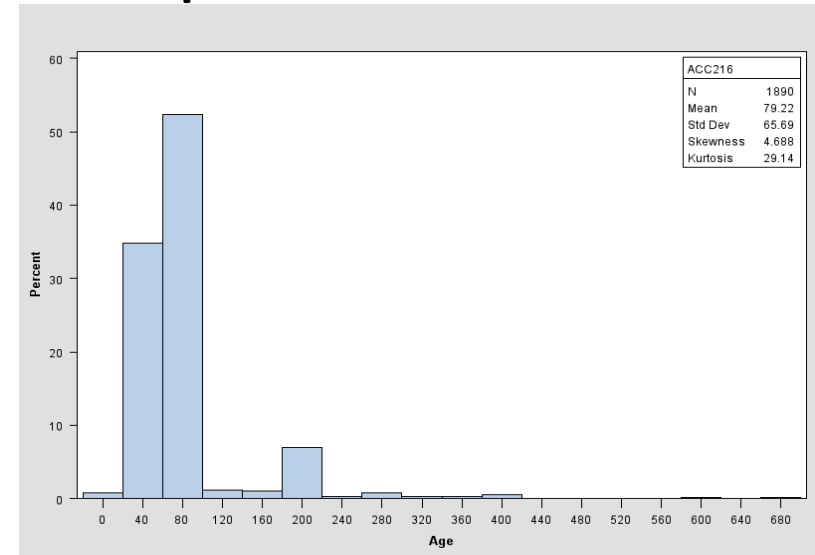
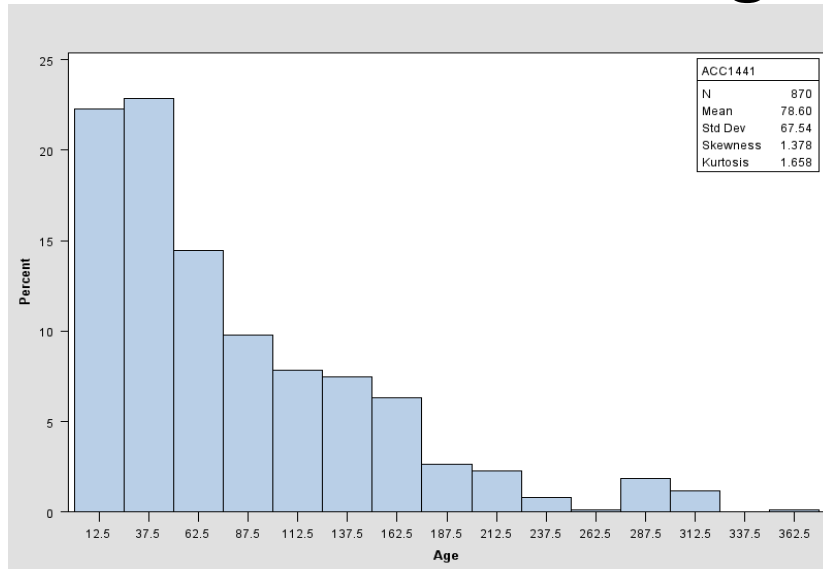
We find that project margin is roughly symmetric with much lower variation. Thus control is probably good in this area. However, RP has a much larger variation and many projects have low RP. **Possibly operational controls are good but sales is not very effective in getting good rates!**

Another Example (Histogram)

Suppose we are studying the payment pattern of two customers. We take the time to pay invoices as the variable of interest. The mean and standard deviations of the two variables are given below. Examine the table carefully and give your comments – do you think the two customers are similar with respect to their payment behaviour?

Parameter	Customer 1	Customer 2
Mean payment time (Days)	78.60	79.22
Standard Deviation (Days)	67.54	65.69

Histogram Example



- The two histograms are very different showing that the payment patterns are not same at all – a fact not detected by the mean and standard deviation
- What are your comments about the payment pattern? Can you suggest any control measures?

Histogram Example (Continued...)

Customer 1

- The histogram given above for invoice payment time is skewed to the right
- Although the average payment time is 78 days (well below the agreed 90 days limit) many invoices take much longer.
- It appears that there is a systemic issue (may be invoices contain errors, may be they are sent without verification of completion of work, may be they are not sent electronically) and focusing on the delinquent invoices may not be of much help.

Customer 2

- The histogram given above shows a different pattern. In this case most invoices are paid within 90 days. However, a few takes much longer, thereby increasing the average time.
- The average time in this case is 79 days – slightly greater than the previous case. However, control is much easier as we probably have to focus on a few specific cases.

Profiling Categorical Variables

- For categorical variables we find the proportion of observations for the different values
- For ordinal variables we can find the median and percentiles
- Suppose the values of the categorical variable are $i = 1, 2, 3 \dots k$. The variance is often computed as $\sum p_i(1 - p_i)$, where p_i gives the relative frequency of i .

Example – Studying a Categorical Variable

Suppose a telecom service provider has carried out a survey to understand the importance the customers attach to various aspects of service. The opinion of the customer has been taken in a 7 point (1 to 7) scale where 1 implies almost no importance and 7 indicates the highest level of importance. The data collected for 357 customers for two service characteristics, namely store experience and consistency of service delivery are tabulated below:

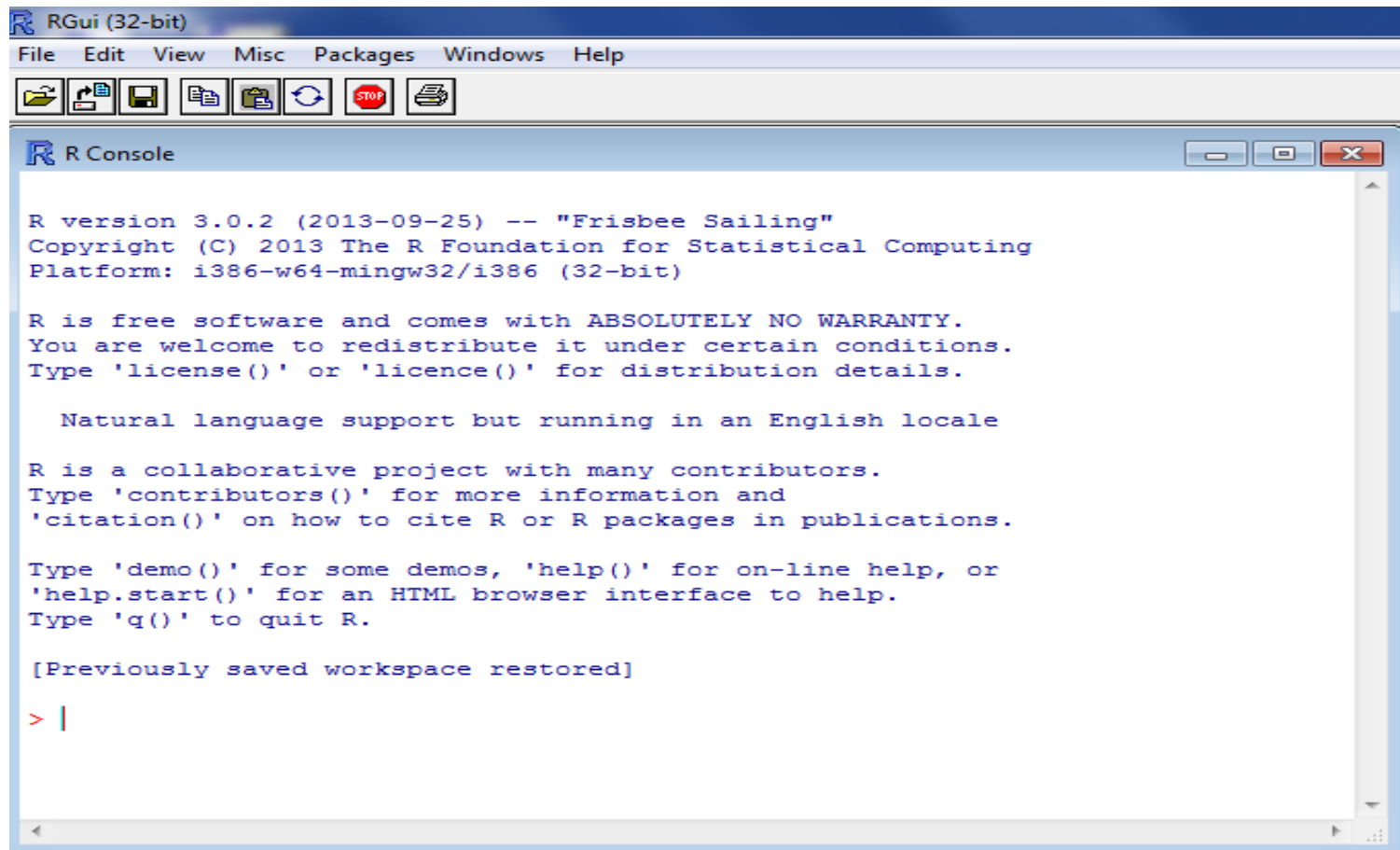
Rating	Variable					
	Store Experience			Consistency of Service Delivery		
	Freq	Prop	Cum Prop	Freq	Prop	Cum Prop
1	4	0.011	0.011	1	0.003	0.003
2	6	0.017	0.028	1	0.003	0.006
3	7	0.020	0.048	5	0.014	0.020
4	35	0.100	0.148	34	0.095	0.115
5	122	0.341	0.489	97	0.272	0.387
6	134	0.375	0.864	155	0.434	0.821
7	49	0.136	1.000	64	0.179	1.000
Total	357	1.000		357	1.000	

Do you see any pattern? Do you understand the profile?

**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



The screenshot shows the RGui (32-bit) application window. The title bar reads "RGui (32-bit)". The menu bar includes "File", "Edit", "View", "Misc", "Packages", "Windows", and "Help". Below the menu bar is a toolbar with icons for file operations and execution. The main window is the "R Console", which displays the following text:

```
R version 3.0.2 (2013-09-25) -- "Frisbee Sailing"
Copyright (C) 2013 The R Foundation for Statistical Computing
Platform: i386-w64-mingw32/i386 (32-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

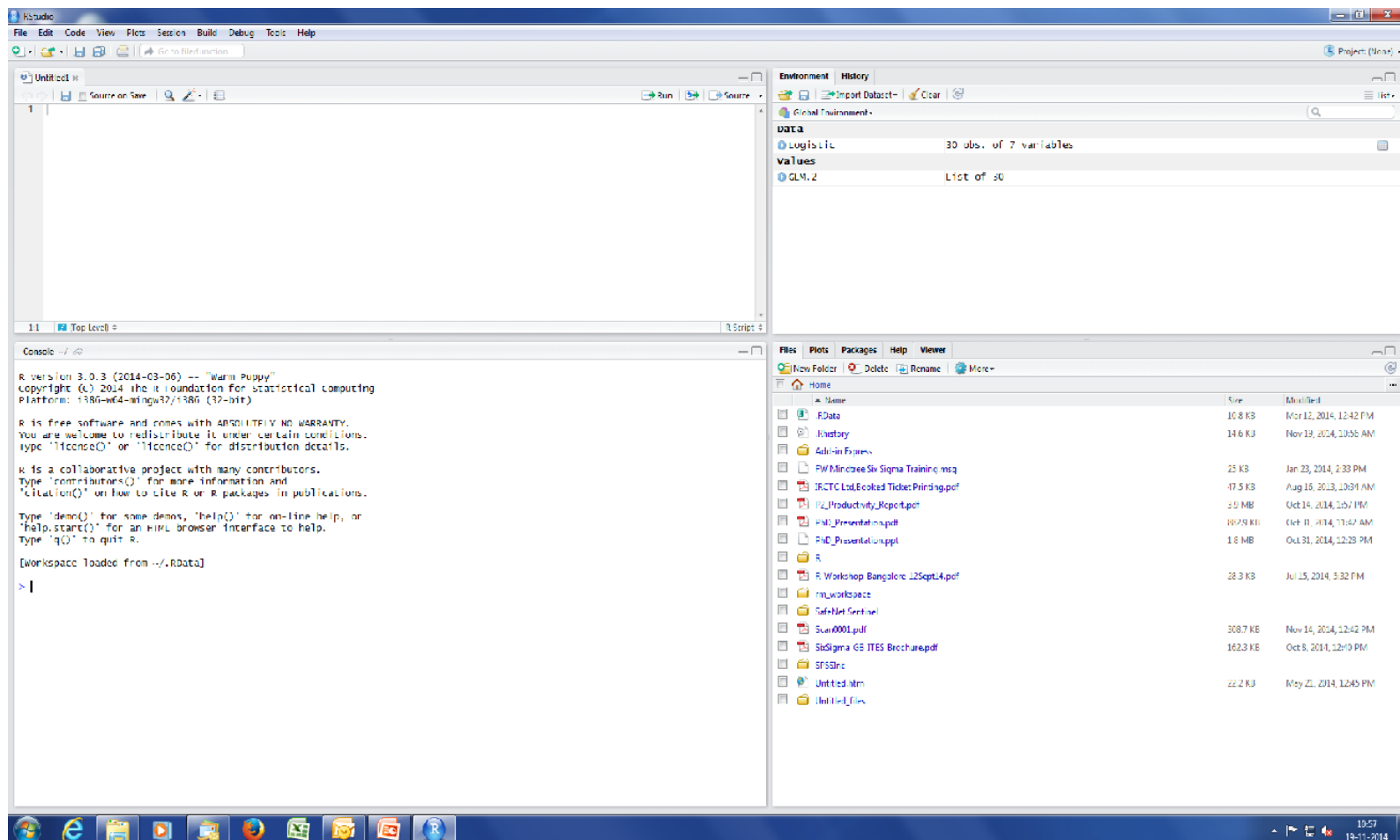
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Previously saved workspace restored]

> |
```

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

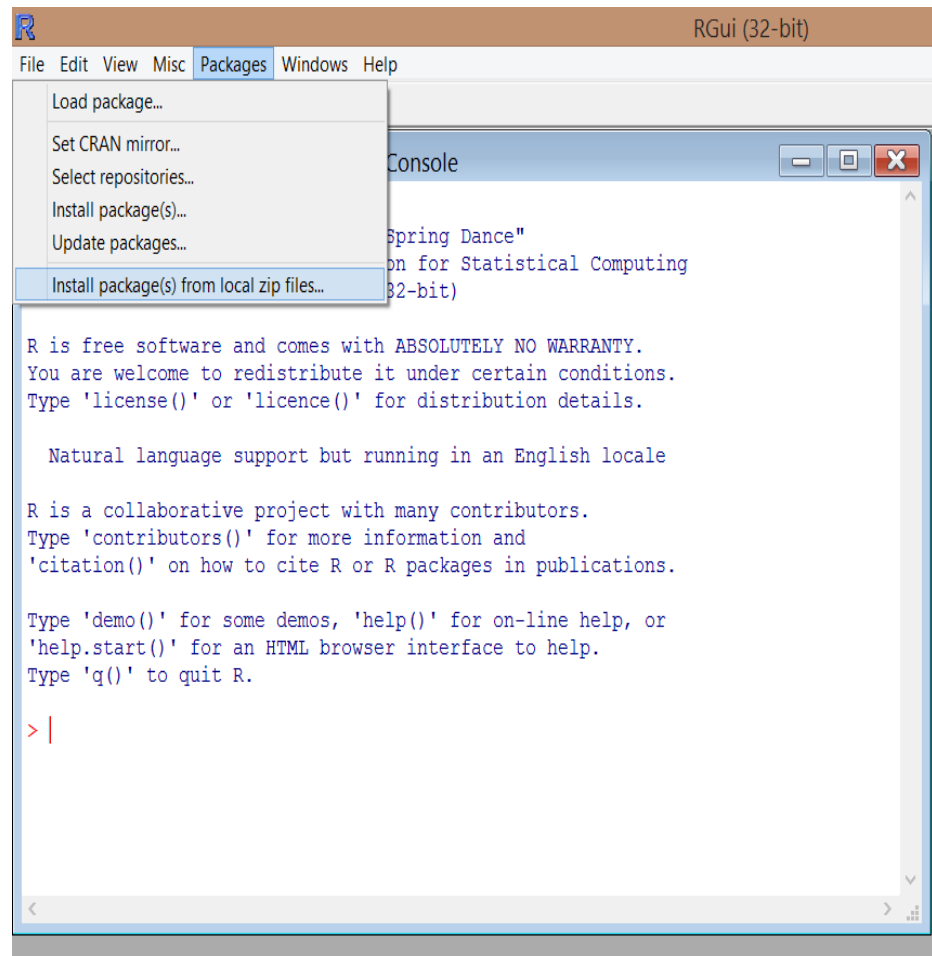


Installing R Libraries

7.1 Download the additional libraries in a folder

7.2 Start R by double clicking R icon

7.3 Choose Install packages(s) from local zip files from Packages menu



DESCRIPTIVE STATISTICS
using R

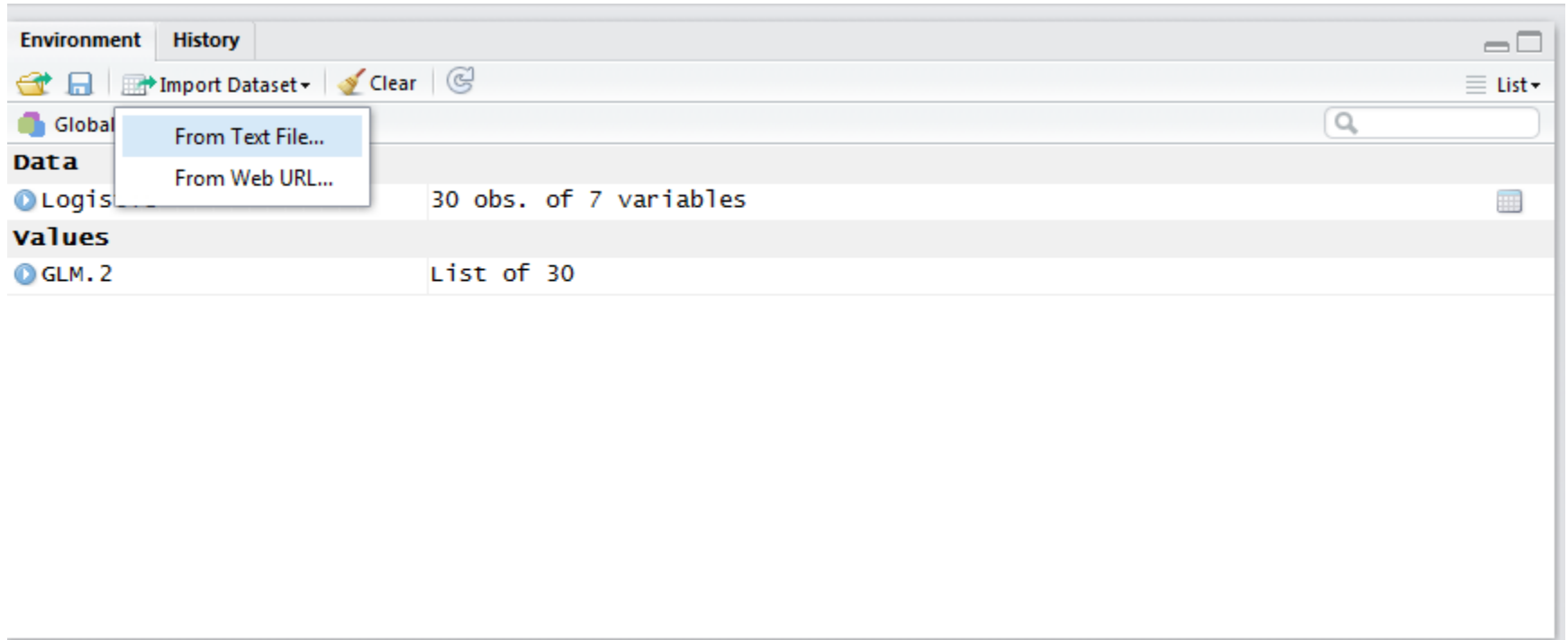
DESCRIPTIVE STATISTICS

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

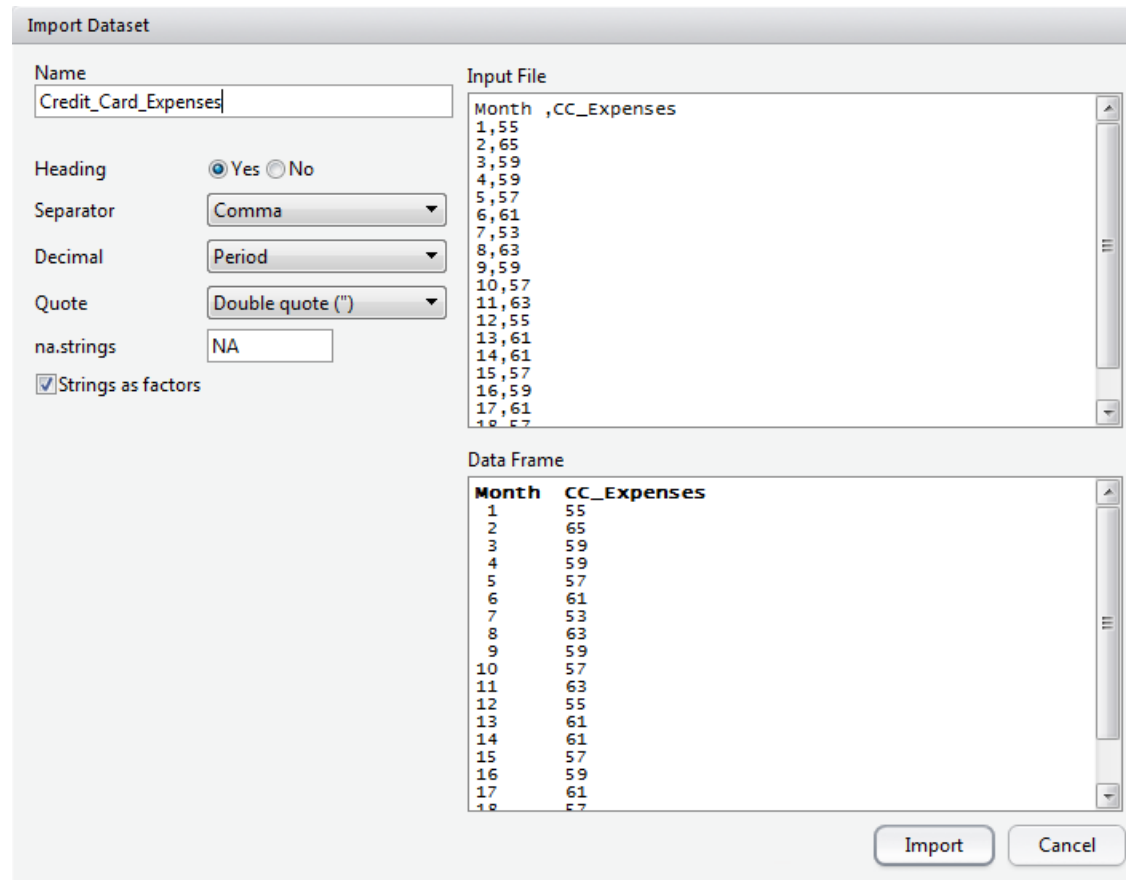
Reading a csv file to R Studio



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	<pre>library(xlsx) mydata <- read.xlsx("c:/myexcel.xlsx", "Sheet1")</pre>

DESCRIPTIVE STATISTICS

Reading data from databases to R Studio

Function	Description
<code>odbcConnect(dsn, uid="", pwd="")</code>	Open a connection to an ODBC database
<code>sqlFetch(channel, sqtable)</code>	Read a table from an ODBC database into a data frame
<code>sqlQuery(channel, query)</code>	Submit a query to an ODBC database and return the results
<code>sqlSave(channel, mydf, tablename = sqtable, append = FALSE)</code>	Write or update (append=True) a data frame to a table in the ODBC database
<code>sqlDrop(channel, sqtable)</code>	Remove a table from the ODBC database
<code>close(channel)</code>	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	<code>> mean(CC)</code>	59.2
Median	<code>> median(CC)</code>	59
Standard deviation	<code>> sd(CC)</code>	3.105174
Variance	<code>> var(CC)</code>	9.642105
Minimum	<code>> min(CC)</code>	53
Maximum	<code>> max(CC)</code>	65
Range	<code>> range(CC)</code>	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

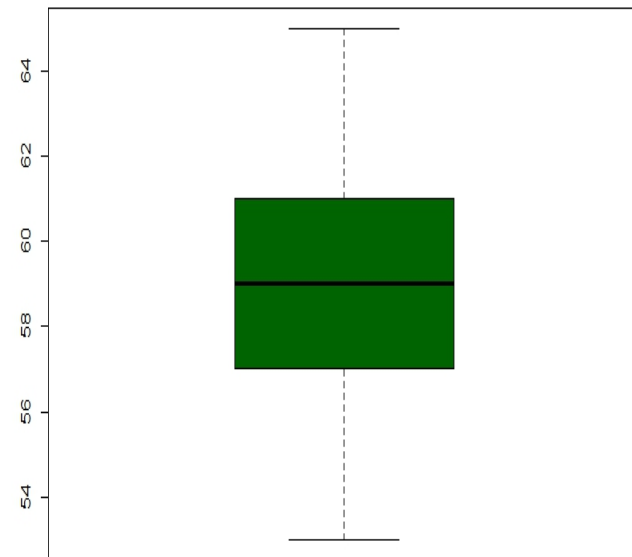
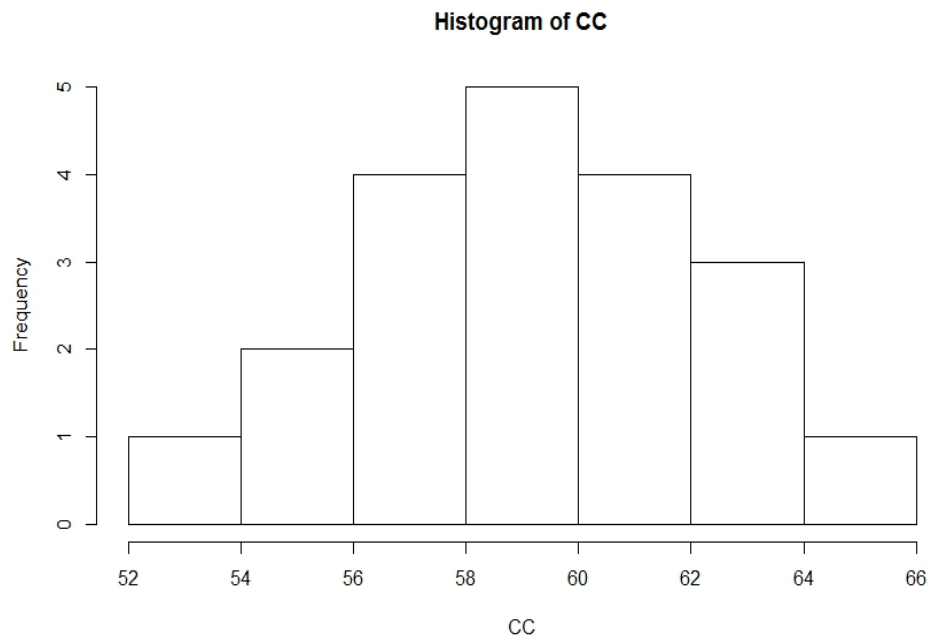
Function	Code
describe	<pre>> library(psych) > describe(CC)</pre>

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	<code>> hist(CC)</code>
Histogram colour ("Blue")	<code>> hist(CC,col="blue")</code>
Dot plot	<code>> dotchart(CC)</code>
Box plot	<code>> boxplot(CC)</code>
Box plot colour	<code>> boxplot(CC, col="dark green")</code>



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)
```

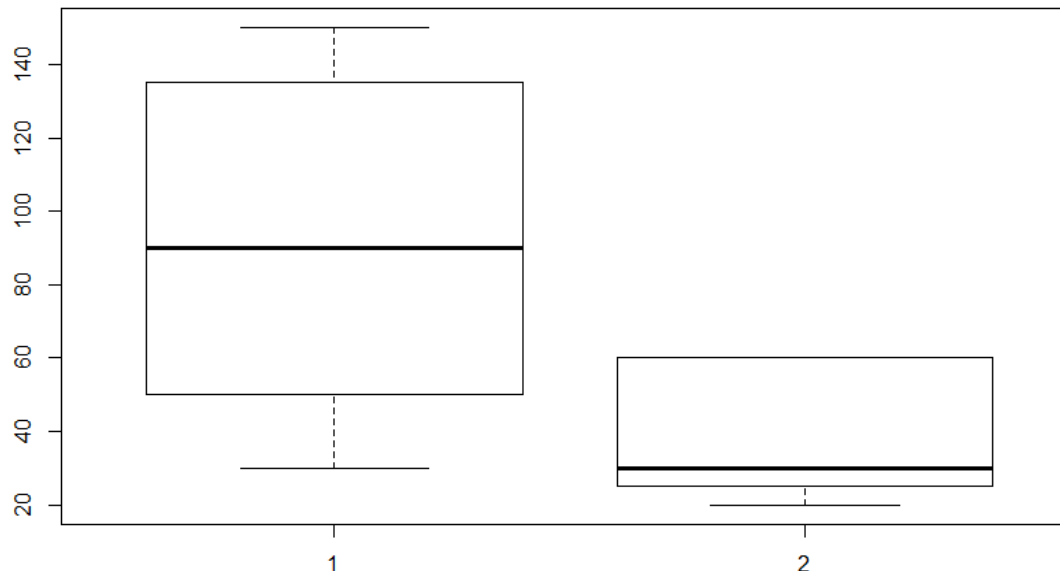
Group	Sex	Average Credit Card Usage
1	Male	93.33333
2	Female	38.66667

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)
```

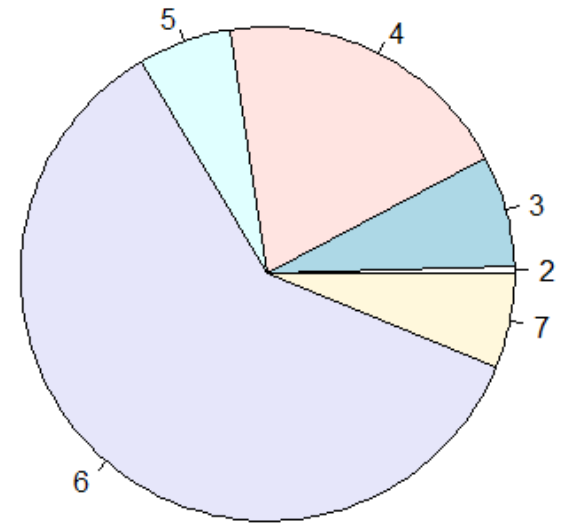
```
> print(mytable)
```

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

DESCRIPTIVE STATISTICS

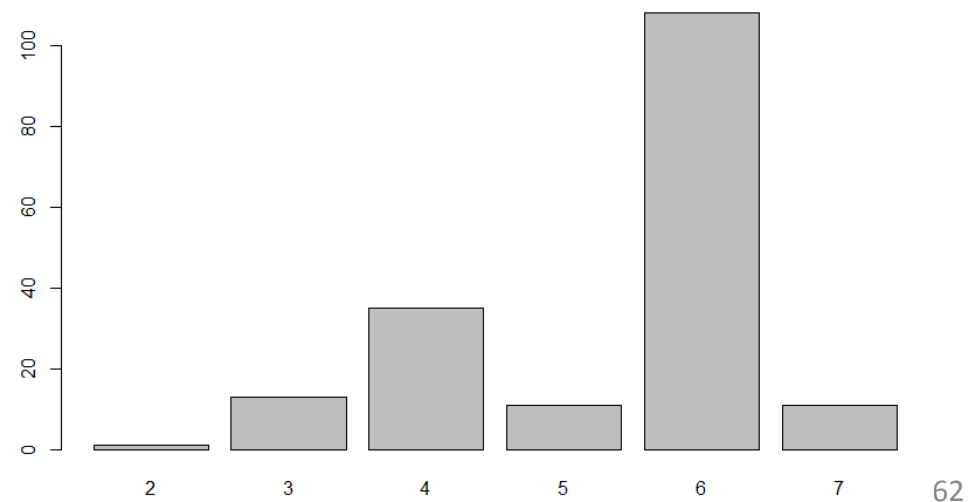
Exercise 3: Creating pie chart for Q4

```
> pie(mytable)
```



Exercise 3: 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 the 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")
```

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

FUNDAMENTALS 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.

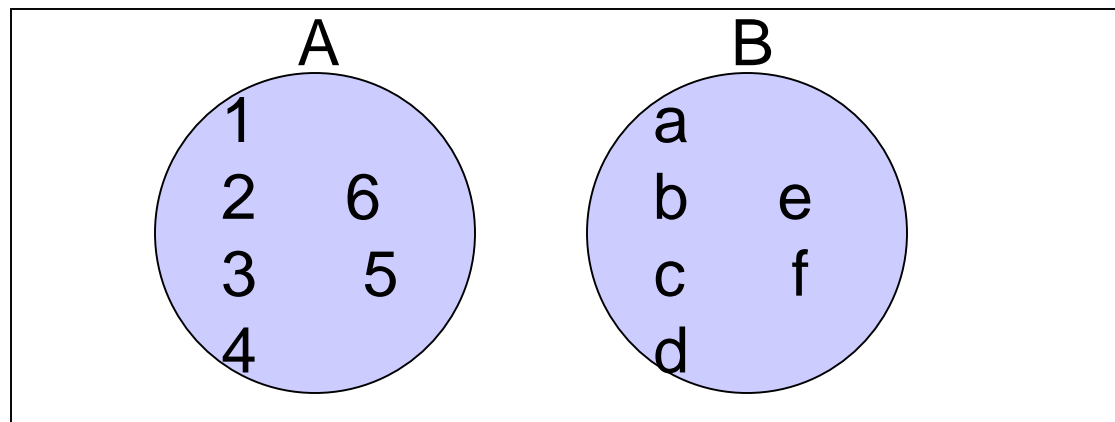
Tossing a coin, Drawing a card from a deck of cards.

FUNDAMENTALS OF PROBABILITY

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}.



Sample space of the diagram - ?

FUNDAMENTALS 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.

FUNDAMENTALS OF PROBABILITY

Example

Tossing of a coin is an experiment.

Here,

Sample Space $S=\{\text{head},\text{tail}\};$

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 $p(\text{getting head})= 1/2$

FUNDAMENTALS OF PROBABILITY

Probability is a number that is assigned to each member of a collection of events from a random experiment that satisfies the following properties:

If S is the sample space and E is any event in a random experiment,

(1) $P(S) = 1$

(2) $0 \leq P(E) \leq 1$

(3) For two events E_1 and E_2 with $E_1 \cap E_2 = \emptyset$

$$P(E_1 \cup E_2) = P(E_1) + P(E_2)$$

Conditional Probability

- Conditional probability of event A given event B – written as $P(A | B)$ is the relative frequency of A given B has happened.
- Conditional probability $P(A | B) = P(AB) / P(B)$. Actually $P(A | B) = N_{AB} / N_B$
- In the table given in the example-cum-exercise slide, what is the conditional probability that a customer will rate his billing experience as 7 given that his experience score is > 5 ?
- Suppose a family has three siblings. What is the conditional probability that the family has three daughters given that out of the 3 siblings at least two are girls?

- $P(A | B)$ is defined only if $B \neq \phi$, i.e. only if $P(B) > 0$
- Note that $P(A | B)$ and $P(B | A)$ are not the same.

An Important Point

- Note that $P(A|B)$ and $P(B|A)$ are not the same. Consider the following example.
- An epidemiologist wants to assess the impact of smoking on the incidence of lung cancer. From hospital records she collected data on 100 patients of lung cancer and she also collected data on 300 persons not suffering from lung cancer. She has classified the 400 samples into smokers and non smokers and the observations are summarized below

Smoker	Lung Cancer		Total
	Yes	No	
Yes	69	137	206
No	31	163	194
Total	100	300	400

- Let A be the event that a person has lung cancer and let B be the event that the person is a smoker. Can you estimate $P(A|B)$ from the table given above?

FUNDAMENTALS 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.

Concept of Independence

- We say that events A and B are independent in case $P(A|B) = P(A)$, i.e. the probability of A is not impacted by the presence of event B
- This definition implies that when A and B are independent, $P(AB) = P(A).P(B)$
- Example: Suppose a machine may fail for three different reasons and suppose these three reasons happen independently. Let A , B and C denote the events that reason 1, reason 2 and reason 3 are present. Then $P(ABC) = P(A).P(B).P(C)$
- **Note:** If A and B are independent, then A^c and B^c are independent. In fact it can be easily shown that A^c and B , and A and B^c are also independent.

Examples of Independence

- Suppose you are tossing a fair coin. Thus the probability that a toss results in a head is 0.5. Assuming that tosses are independent of each other, what is the chance that 3 tosses will result in 3 heads?
- Suppose a machine has 20 different parts. Suppose the parts fail independently of each other and on any given day a part fails with 1% chance only. Suppose the machine continues to operate if all parts are operational and fails if one or more parts fail. What is the chance that the machine will fail on a randomly selected day?

Random Experiments and Random Variables

- Certain variables are generated as some activities are carried out. A flight takes certain amount of time; a machine when operated consumes certain amount of power per unit time; a car travels certain number of miles before encountering a fault. The activities giving rise to the data are called *random experiments*. These are the *data generating mechanisms*. The generated data are often referred to as *random variables*.
- Technically random variables are mappings that assigns a real number to every outcome of the activity. We often write this as $X:\Omega \longrightarrow \mathbb{R}$, where Ω is the sample space containing all possible outcomes of the random experiments

Random Variables

- Random variables take values from a predefined set. Thus all possible values of the random variable are known in advance. However, the exact value that will occur in the next occurrence is unknown.
- We do not know when a machine will fail, what will be power consumption of a device in a given period, what will be the efficiency of a machine, how many defective products will be produced during the next production run...

Concept of Distributions

- It is often assumed that the process that generates the random variable can be modeled using a mathematical function. This function enables us to calculate the probability that the values of a random variable will fall within any given range
- The mathematical function is defined in terms of the random variable and some unknown but fixed constants called parameters
- The distribution (mathematical function) characterizes the pattern of variation of the random variable
- We may construct the frequency distribution as well to visualize the pattern of variation. When we have very large number of observations drawn randomly from a population, the frequency distribution may be a good approximation of the distribution.
- The frequency distribution is a non-parametric way of looking at a distribution and it has its own advantages and disadvantages.

FUNDAMENTALS OF PROBABILITY

Binomial Distribution

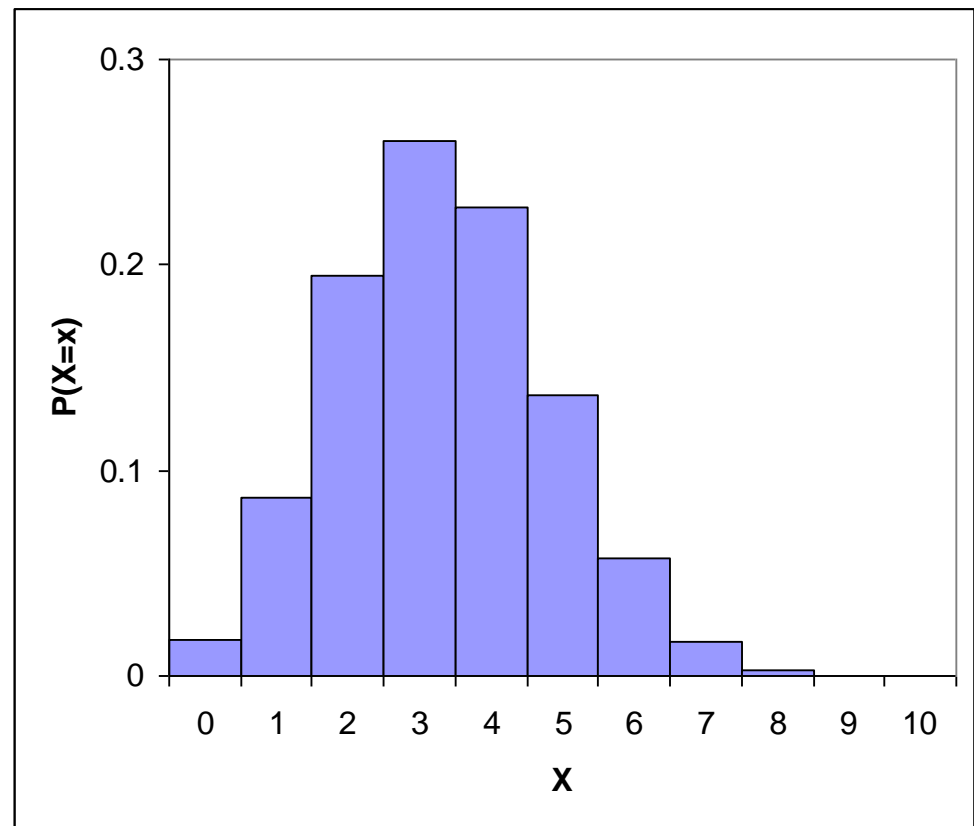
The number of successes in n Bernoulli trials.

Or the sum of n Bernoulli random variables.

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

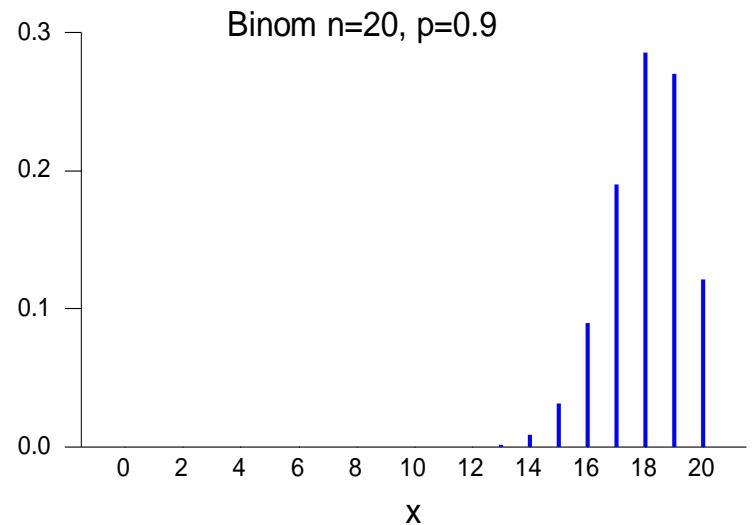
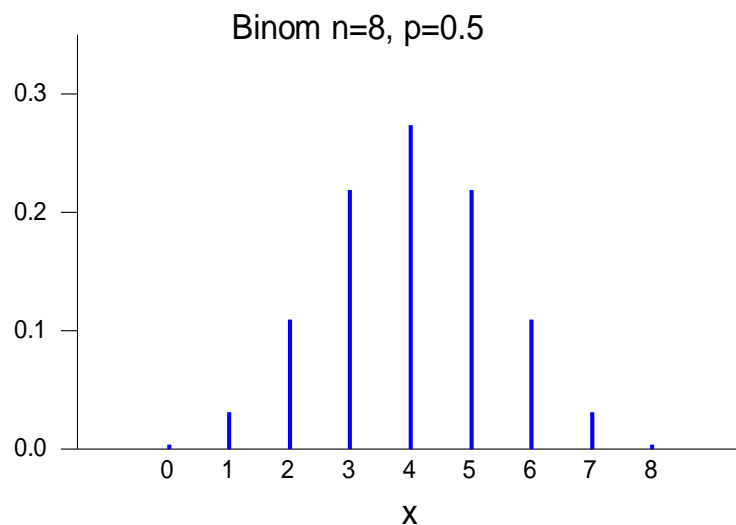
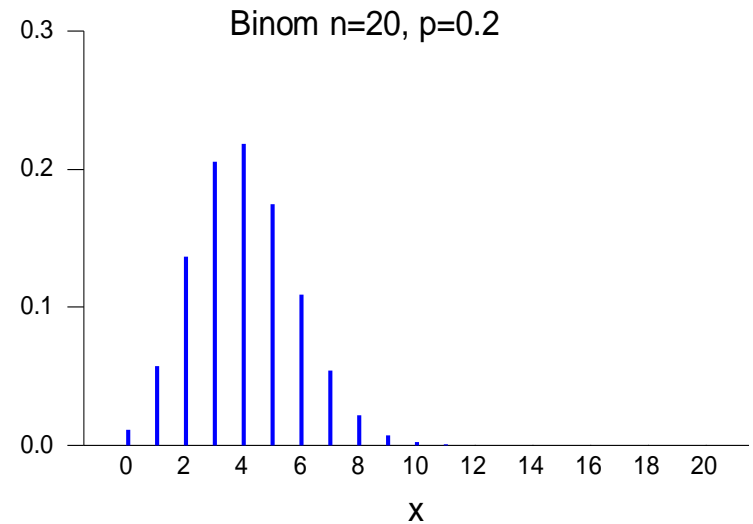
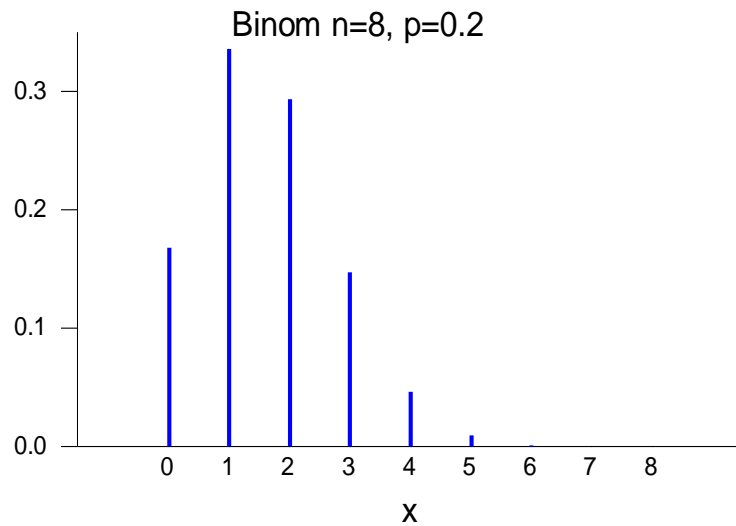
$$E[X] = np$$

$$\text{Var}(X) = np(1-p)$$



FUNDAMENTALS OF PROBABILITY

Binomial Distribution Plots



FUNDAMENTALS OF PROBABILITY

Poisson Distribution

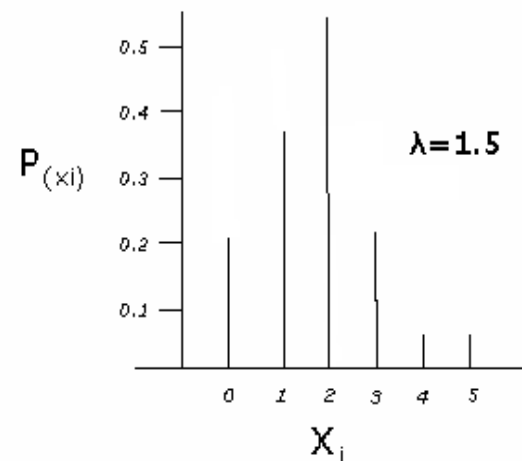
- Poisson distribution also describes discrete data – situations where the random variable can take integer values.

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 =

$$\sqrt{\text{mean}}$$

Poisson Distribution

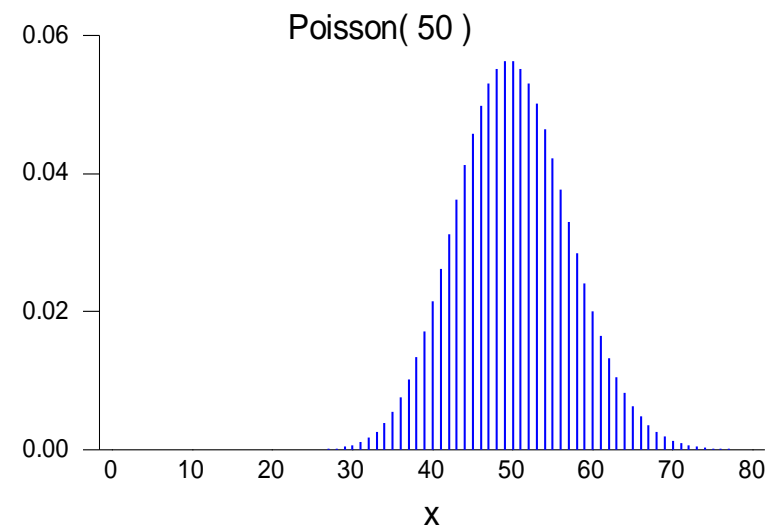
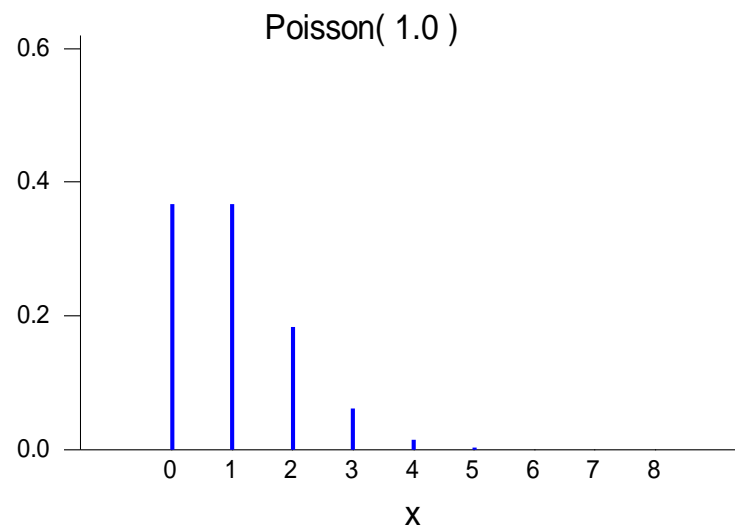
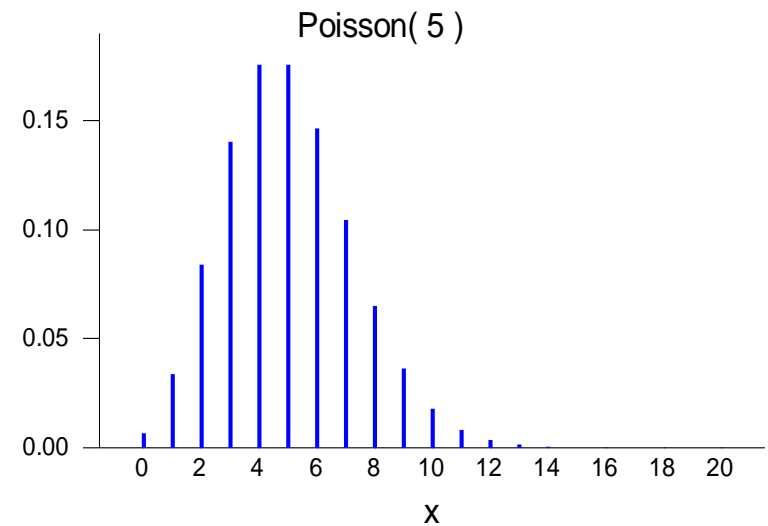
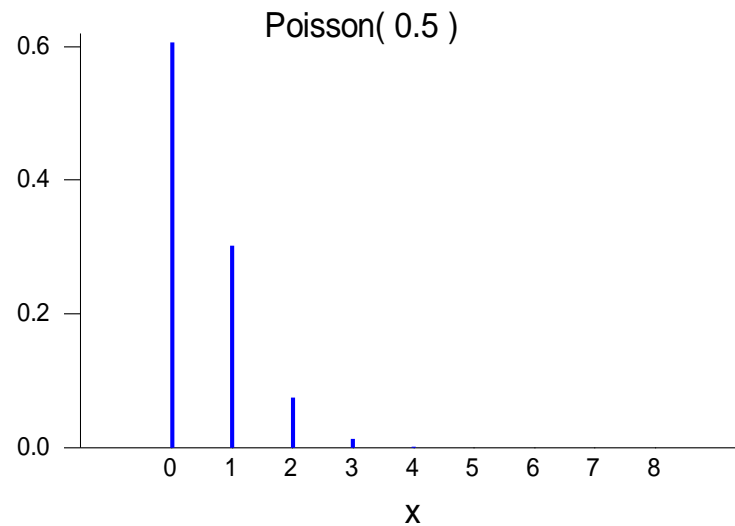


$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

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.

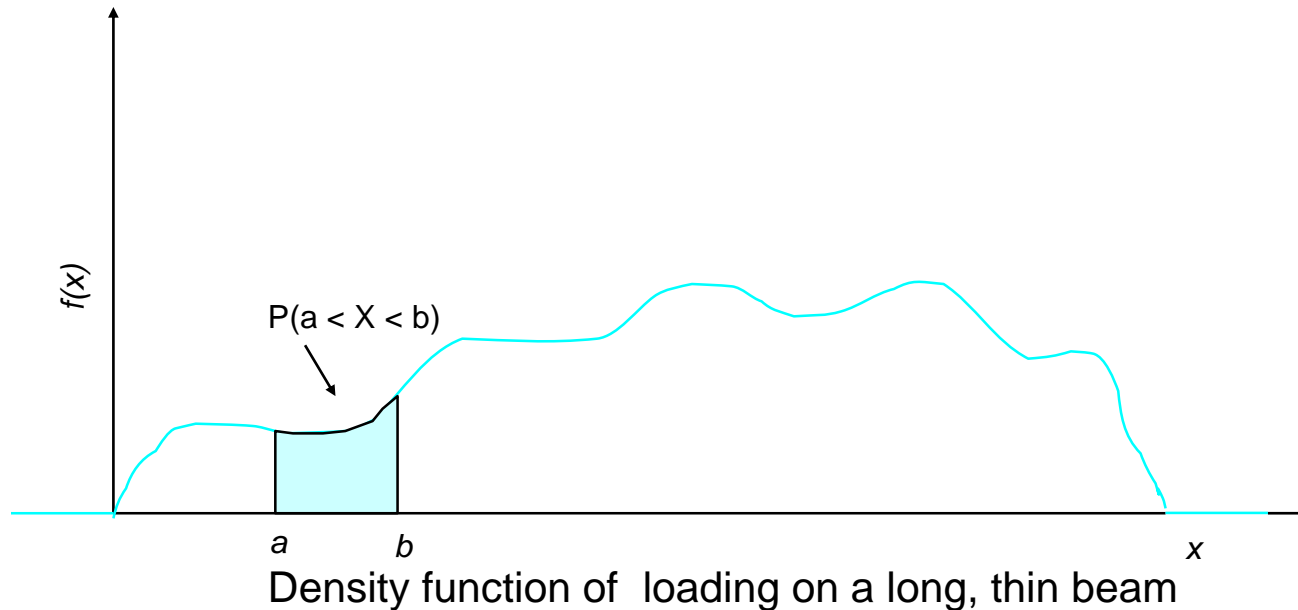
FUNDAMENTALS OF PROBABILITY

Poisson Distribution Plots



FUNDAMENTALS OF PROBABILITY

Probability Density Function



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

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

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

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

FUNDAMENTALS OF PROBABILITY

Uniform Distribution

A continuous random variable X with probability density function

$$f(x) = 1/(b-a), \quad a \leq x \leq b$$

has a **continuous uniform distribution**

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

$$\mu = E(X) = (a+b)/2 \quad \text{and} \quad \sigma^2 = V(X) = (b-a)^2/12$$

Applications:

- Generating random sample
- Generating random variable

FUNDAMENTALS OF PROBABILITY

Normal Distribution

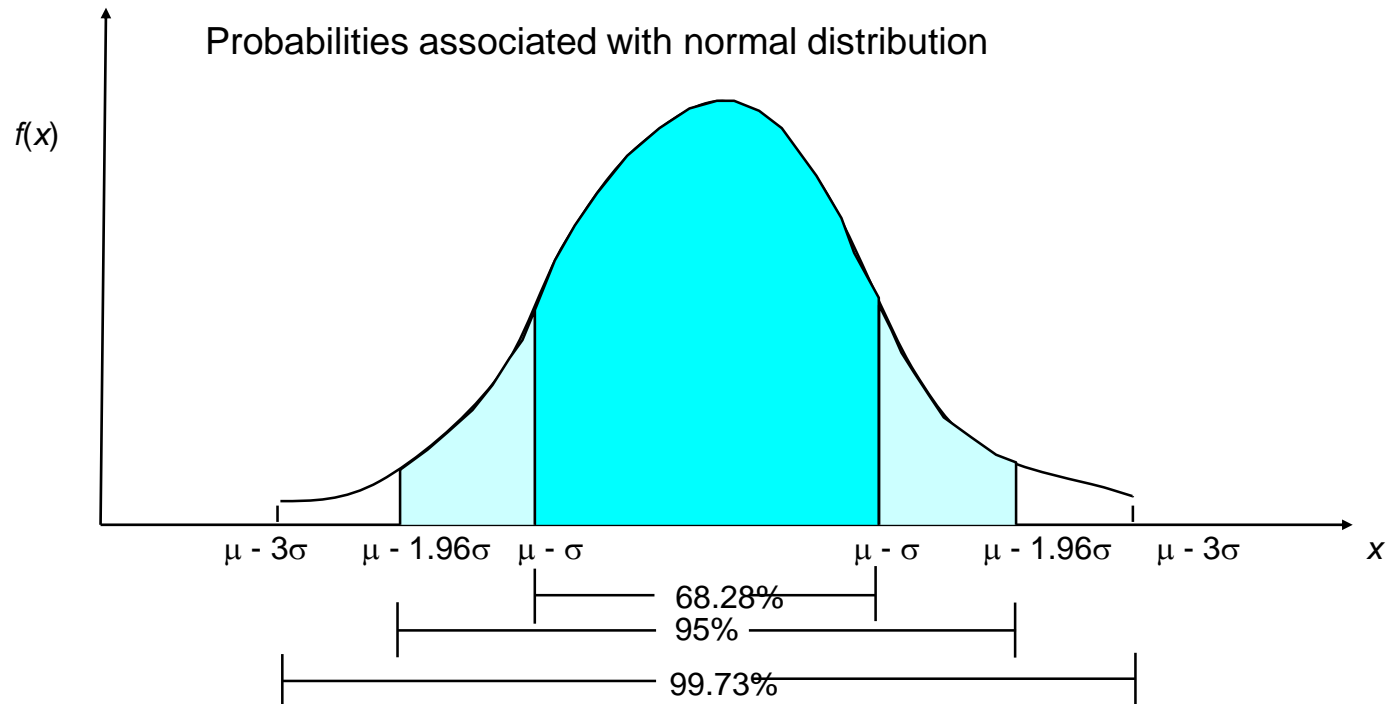
A random variable X with probability density function

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

has a normal distribution with parameters μ , where $-\infty < \mu < \infty$, and $\sigma > 0$. Also,

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

Probabilities associated with normal distribution



FUNDAMENTALS OF PROBABILITY

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 .

The CDF of a standard normal random variable is denoted as

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

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.

FUNDAMENTALS OF PROBABILITY

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

FUNDAMENTALS OF PROBABILITY

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

$$f(x) = \lambda e^{-\lambda x}, \quad \text{for } 0 \leq x < \infty$$

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

$$E(X) = 1/\lambda \quad \text{and} \quad V(X) = 1/\lambda^2$$

FUNDAMENTALS OF PROBABILITY

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

DISTRIBUTIONS

R Functions

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

DISTRIBUTIONS

Binomial Distribution

Exercise 1: An electronic product contains 40 integrated 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?

DISTRIBUTIONS

Binomial Distribution

Exercise 1: An electronic product contains 40 integrated 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

DISTRIBUTIONS

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?

DISTRIBUTIONS

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.

- What is the probability that exactly 5 tickets arrives in one hour?
- What is the probability that 3 or less tickets arrives in one hour?
- What is the probability that 15 or more tickets arrives in two hour?
- 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

DISTRIBUTIONS

Normal Distribution

- Exercise 1:** The compressive strength of samples of cement can be modelled by a normal distribution with mean of 6000 kg/cm² and a standard deviation of 100 kg/cm².
- What is the probability that a sample's strength is less than 6250 kg/cm²?
 - What is the probability that a sample's strength is between 5800 and 5900 kg/cm²?
 - What strength is exceeded by 95% of the samples?

DISTRIBUTIONS

Normal Distribution

Exercise 1: The compressive strength of samples of cement can be modelled by a normal distribution with mean of 6000 kg/cm² and a standard deviation of 100 kg/cm².

- What is the probability that a sample's strength is less than 6250 kg/cm²?
- What is the probability that a sample's strength is between 5800 and 5900 kg/cm²?
- 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/cm² = 0.99379

DISTRIBUTIONS

Normal Distribution

- Exercise 2:** The tensile strength of a paper is modelled by a normal distribution with mean of 35 pounds/inch² and a standard deviation of 2 pounds/inch².
- What is the probability that a sample's strength is less than 40 pounds/inch²?
 - If the specification of tensile strength is not to exceed 35pounds/inch², what proportion of the samples is scrapped?

DISTRIBUTIONS

Normal Distribution

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?

DISTRIBUTIONS

Exponential Distribution

- Exercise 1:** The time to failure (t hours) for a laser in a cytometry machine is modelled by an exponential distribution with $\lambda = 0.00004$?
- What is the probability that the laser will not fail in 20000 hours?
 - What is the probability that the laser will not last 30000 hours?

DISTRIBUTIONS

Exponential Distribution

- Exercise 1:** The time to failure (t hours) for a laser in a cytometry machine is modelled by an exponential distribution with $\lambda = 0.00004$?
- What is the probability that the laser will not fail in 20000 hours?
 - 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

DISTRIBUTIONS

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)
```


Summary of the Distributions

Name	Parameters	Mean	Variance
Binomial	p	np	$np(1 - p)$
Poisson	λ	λ	λ
Geometric	p	$1 / p$	$(1 - p) / p^2$
Negative Binomial	p	$[r(1 - p)] / p$	$r(1 - p) / p^2$
Discrete Uniform	K	$(K + 1) / 2$	$(K - 1)(K + 1) / 12$
Uniform	a, b	$(a + b) / 2$	$(b - a)^2 / 12$
Normal	μ, σ^2	μ	σ^2
Exponential	λ	$1 / \lambda$	$1 / \lambda^2$
Weibull	α, β	$\beta \Gamma(1 + 1/\alpha)$	$\beta^2 \{ \Gamma(1 + 2/\alpha) - (\Gamma(1 + 1/\alpha))^2 \}$

Usages of Distributions

- Distributions of random variables may be used from three perspectives, namely to understand
 - To find the probability of any event
 - To compare two or more distributions; or to compare effectiveness of some actions
 - To simulate real life scenarios using the distributional models to develop an understanding of the phenomenon

TEST
of
HYPOTHESIS

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
 - A ten-percent increase in price will not adversely affect the sale of this product

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 H_0

Alternative Hypothesis:

One in which some difference or effect is expected

Denoted by H_1

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 H_0 when it is true

Type II Error

Failing to reject the null hypothesis H_0 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 or a phenomenon needs explanation
2. Formulate the model (Create the Statistical Problem: $Y = f(X)$)
3. Establish the Hypotheses
 - State the Null Hypothesis (H_0)
 - State the Alternative Hypothesis (H_a).
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
7. Calculate the test statistic (z , t , or F) from the data.
8. Determine the probability of that calculated test statistic occurring by chance.
9. If that probability is less than α , reject H_0 otherwise do not reject H_0 .
10. Replicate results and translate statistical conclusion to practical solution.

TEST OF HYPOTHESIS

Test of Comparisons (X=Discrete)

	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, $\bar{x} = 5.15$

Compare \bar{x} with specified value 5

or $\bar{x} - \text{specified value} = \bar{x} - 5$ with 0

If $\bar{x} - 5$ 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, $\bar{x} = 515$

$$\bar{x} - 500 = 515 - 500 = 15$$

Can we conclude $\mu \neq 500$?

Conclusion:

Difficult to say $\mu = \text{specified value}$ by looking at $\bar{x} - \text{specified value}$ alone

TEST OF HYPOTHESIS

Methodology demo: To Test $\mu = \text{Specified Value}$ ($\mu = \mu_0$)

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

To test Mean = Specified value

$$\text{Test Statistic } t_0 = (\text{xbar} - \text{Specified value}) / (\text{SD} / \sqrt{n})$$

If **test statistic** is close to **0**, conclude that $\mu = \text{Specified value}$

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

TEST OF HYPOTHESIS

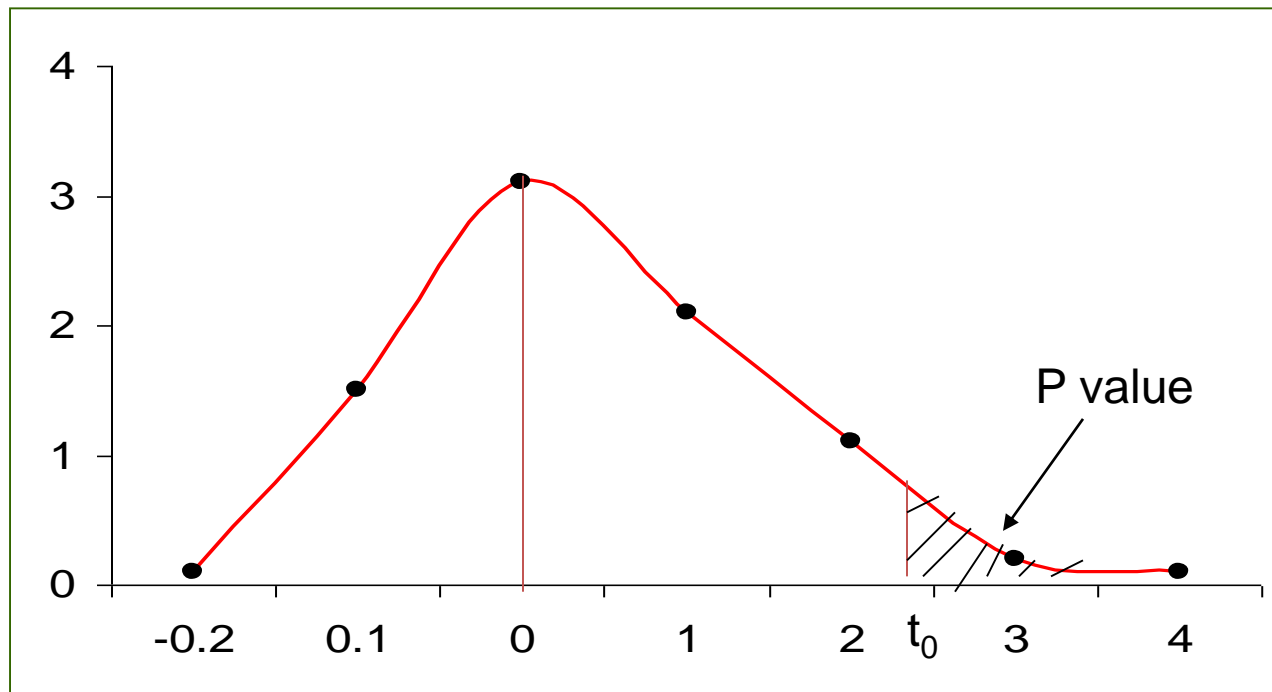
Methodology demo: To Test $\mu = \text{Specified Value}$

P value

The probability that such evidence or result will occur when H_0 is true

Based on the reference distribution of test statistic

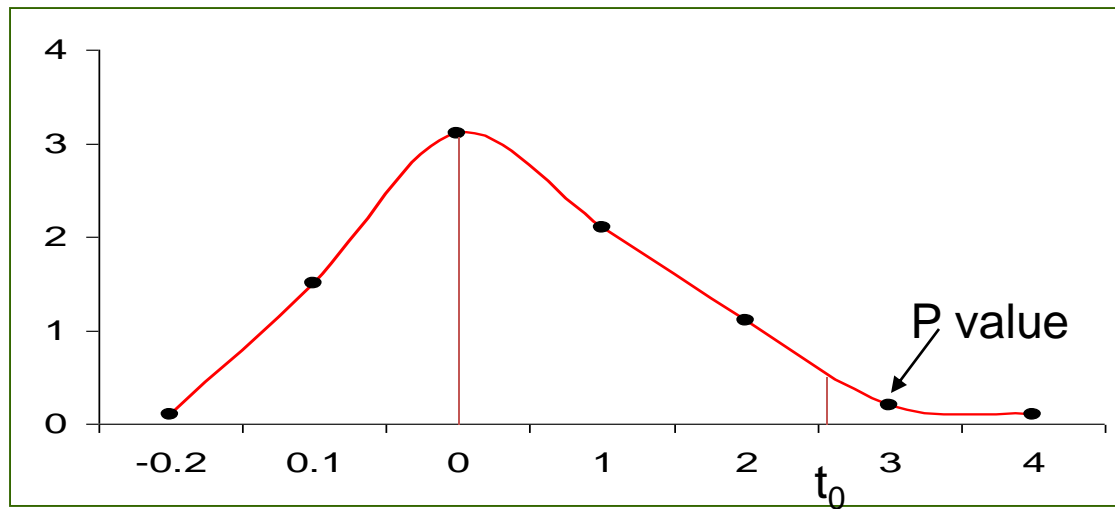
The tail area beyond the value of test statistic in reference distribution



TEST OF HYPOTHESIS

Methodology demo : To Test $\mu = \text{Specified Value}$

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

$\mu \neq \text{Specified Value}$, H_0 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

$$H_0: \mu = 5$$

$$H_1: \mu \neq 5$$

Calculate $\bar{x} = 5.15$

$$SD = 0.8515$$

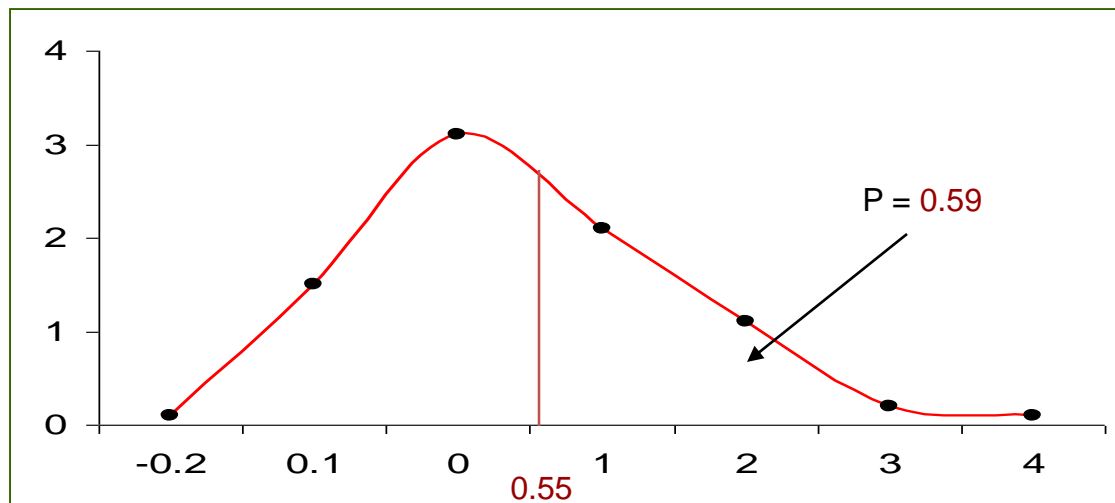
$$n = 10$$

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

TEST OF HYPOTHESIS

Example: To Test $\mu = \text{Specified Value } (\mu = \mu_0)$

$$t_0 = 0.5571$$



$P \geq 0.05$, hence $\mu = \text{Specified value} = 5$.

H_0 : Mean = 5 is not rejected

TEST OF HYPOTHESIS

Hypothesis Testing: Steps

1. Formulate the null hypothesis H_0 and the alternative hypothesis H_1
2. Select an appropriate statistical test and the corresponding test statistic
3. Choose level of significance α (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 Complaint_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 \bar{x}_1 & \bar{x}_2

Calculate SD1 & SD2

Compare \bar{x}_1 with \bar{x}_2

Or $\bar{x}_1 - \bar{x}_2$ with 0

Calculate test statistic t_0 by dividing $(\bar{x}_1 - \bar{x}_2)$ by a function of SD1 & SD2

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

Calculate p value from t distribution

If $p \geq 0.05$ then $H_0: \text{Mean}_1 = \text{Mean}_2$ 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

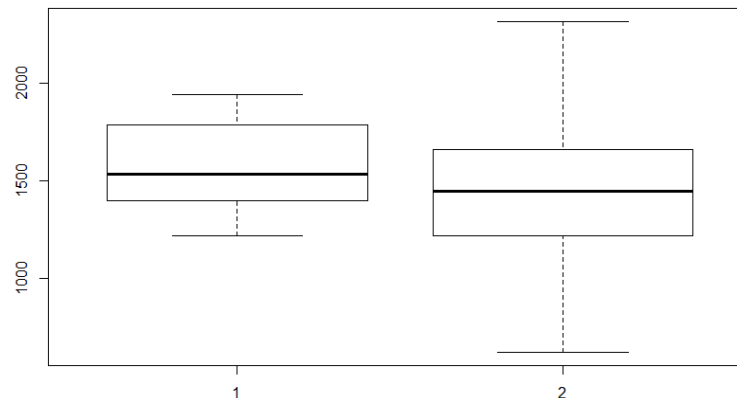
TEST OF HYPOTHESIS

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

$$H_0: \sigma_1^2 = \sigma_2^2$$

Alternative hypothesis

$$H_1: \sigma_1^2 \neq \sigma_2^2$$

Calculate standard deviations of both the samples S1 & S2

Calculate test statistic $F = S_1^2 / S_2^2$

If F is close to 1, then S_1^2 more or less equal to S_2^2

Calculate p from F distribution.

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

$H_0: \sigma_1^2 = \sigma_2^2$ 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 $t_0 = \frac{\bar{d}}{s_d / \sqrt{n}}$

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)
```

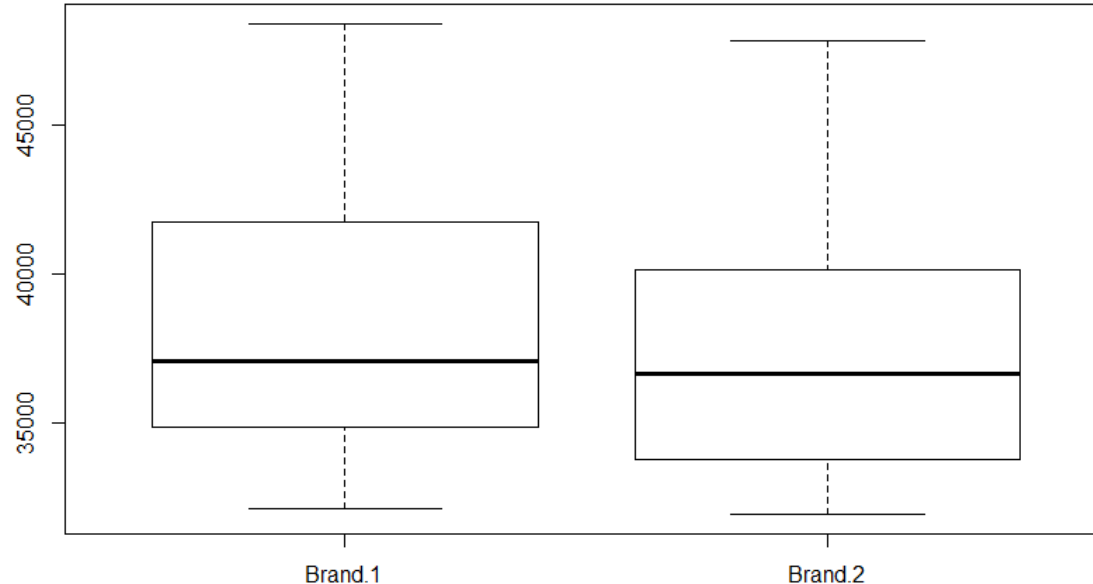
Statistics	Value
t	1.9039
df	7
P value	0.09863

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



TEST OF HYPOTHESIS

Paired t test: Exercise 2

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

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:** $p = \text{Specified Value}$ ($p = p_0$)

Alternative hypothesis **H1:** $p \neq \text{Specified Value}$ ($p \neq p_0$)

or

H1: $p > \text{Specified Value}$ ($p > p_0$)

or

H1: $p < \text{Specified Value}$ ($p < p_0$)

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

$$z_0 = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

Calculate p value from z distribution

If p value ≥ 0.05 then $H_0: p = \text{Specified Value}$ 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 $H_0: p_1 = p_2$

Alternative Hypothesis $H_1: p_1 \neq p_2$

or

$H_1: p_1 > p_2$

or

$H_1: p_1 < p_2$

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 $H_0: p_1 = p_2$ 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

H_0 : Deviation from bell shape (normality) = 0

H_1 : Deviation from bell shape $\neq 0$

If $p \text{ value} \geq 0.05$ (5%), then H_0 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

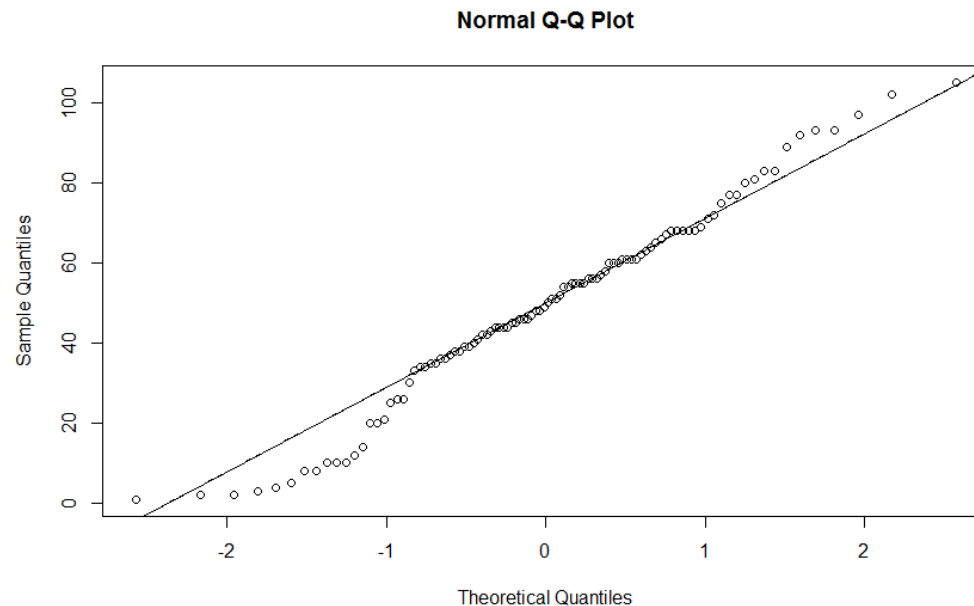
Normality test

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 `Complaint_Response_Time.csv`. Check whether the complaint response time follows normal distribution?

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?

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

**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

$H_0 = \text{Mean}_1 = \text{Mean}_2 = \dots = \text{Mean}_k$

Reject H_0 if $p\text{-value} < 0.05$

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 revenue

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)

$$\begin{aligned} T &= \text{Sum of all the response values} \\ &= 1.55 + 2.36 + \dots + 2.72 + 2.07 = 53.33 \end{aligned}$$

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

$$N = 18$$

Step 4: Calculate the Correction Factor (CF)

$$\begin{aligned} CF &= (\text{Grand Total})^2 / \text{Number of Response values} \\ &= T^2 / N = (53.33)^2 / 18 = 158.0049 \end{aligned}$$

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

$$\begin{aligned} TSS &= \text{Sum of square of all the response values} - CF \\ &= 1.55^2 + 2.36^2 + \dots + 2.72^2 + 2.07^2 - 158.0049 \\ &= 15.2182 \end{aligned}$$

ANALYSIS OF VARIANCE

One Way ANOVA : Example

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

$$\begin{aligned} SS_A &= A_1^2 / nA_1 + A_2^2 / nA_2 + A_3^2 / nA_3 - CF \\ &= 7.47^2 / 4 + 30.31^2 / 8 + 15.55^2 / 4 - 158.0049 = 11.0827 \end{aligned}$$

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

$$\begin{aligned} SS_e &= \text{Total sum of square} - \text{between sum of square} \\ &= TSS - SS_A = 15.2182 - 11.0827 = 4.1354 \end{aligned}$$

Step 8: Calculate degrees of freedom (df)

$$\text{Total df} = \text{Total Number of response values} - 1 = 18 - 1 = 17$$

$$\text{Between df} = \text{Number of levels of the factor} - 1 = 3 - 1 = 2$$

$$\text{Within df} = \text{Total df} - \text{Between df} = 17 - 2 = 15$$

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				

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

$$F \text{ Crit} = \text{finv}(\text{probability}, \text{between df}, \text{within df}), \text{probability} = 0.05$$

$$P \text{ value} = \text{fdist}(F, \text{between df}, \text{within df})$$

One Way ANOVA : Decision Rule

If $p \text{ value} < 0.05$, then the factor has significant effect on the process output or response. In this example as $p \text{ value} < 0.05$ 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)
```

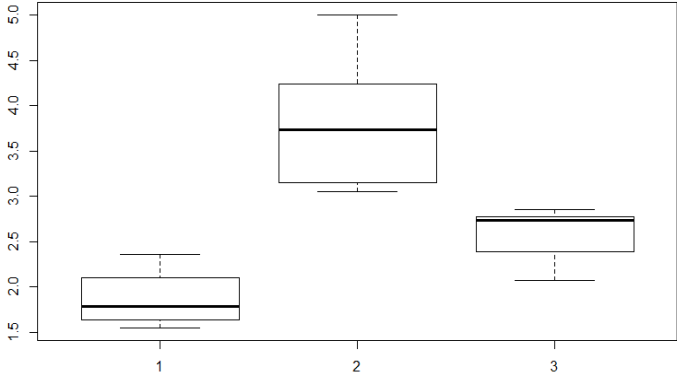
One Way ANOVA : Example Result

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

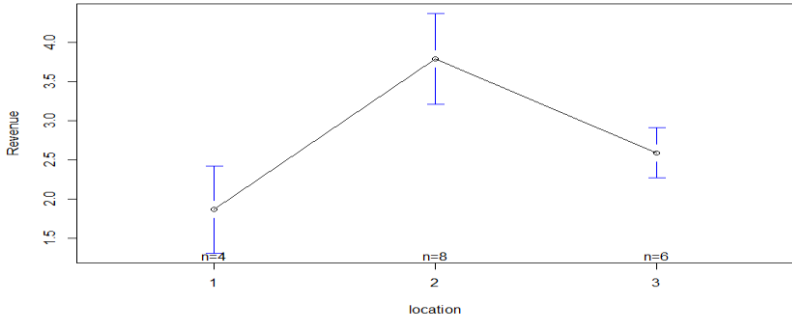
```
> aggregate(Revenue ~ location, FUN = mean)
```

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	3.20	2.30
	1.89	2.81	1.91
	1.35	4.52	1.40
	2.07	4.40	1.48
	2.41	4.75	
	3.06	5.19	
		3.42	
		9.80	

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 $F = MS_{\text{between}} / MS_{\text{within}}$

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 $p \text{ value} \geq 0.05$, then variation within the levels are equal, otherwise not

R Code for Bartlett's test

```
> bartlett.test(Revenue, location, 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 $p \text{ value} = 0.1472 > 0.05$, 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?

Analysis of Variance – Two Way

- **Two-way ANOVA is a type of study design with continuous output variable and two categorical explanatory variables.**
- **Example**
 - **In a study we may wish to study the effect of assembly method and operator on assembly time.**
 - **The effect of capsule types and digestive fluids on dissolving time of capsule.**
 - **The effect of alcohol and gender on attractiveness.**
- **Sometimes we need to study the effect of one factor by blocking the effect of another factor. The factor, which is we are not interested, is defined as a group.**
- **Technical term for such a group is block and the study design is also called randomised block design**
- **Example: The effect of catalyst on yield by blocking the effect of batch.**

Analysis of Variance – Two Way

Example of Blocking:

- An engineer wants to test 4 catalysts. For time reasons, he can only run 4 tests per batch.
- In what sequence do you perform the experiment and why?

	Cat 1	Cat 2	Cat 3	Cat 4
Charge 1	69	72	73	75
Charge 2	72	75	75	74
Charge 3	68	67	68	72
Charge 4	71	72	72	75

Mathematical model

$$y_{ti} = \mu + \beta_i + \tau_t + \varepsilon_{ti}$$

$$\text{Ho: } \beta_i \text{'s} = 0$$

$$\text{Ha: } \beta_i \text{'s} \neq 0$$

$$\text{Ho: } \tau_t \text{'s} = 0$$

$$\text{Ha: } \tau_t \text{'s} \neq 0$$

Open the datafile : yield

Analysis of Variance – Two Way

Example of Blocking:

```
> Catalyst=factor(Catalyst)
> fit=aov(Yield~Catalyst)
➤ summary(fit)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Catalyst	3	32.75	10.92	1.747	0.211
Residuals	12	75.00	6.25		

Catalyst does not effect yield as the p-value > 0.05. However the one way ANOVA is wrong way analyzing the data. We need to eliminate the effect of batch and then study the effect of catalyst.

```
> fit1=aov(Yield~Catalyst+Batch)
summary(fit1)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Catalyst	3	32.75	10.92	6.238	0.0140	*
Batch	3	59.25	19.75	11.286	0.0021	**
Residuals	9	15.75	1.75			

--- signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Catalyst effects yield significantly as the p-value < 0.05.

Analysis of Variance – Two Way

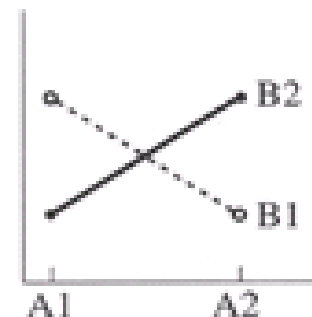
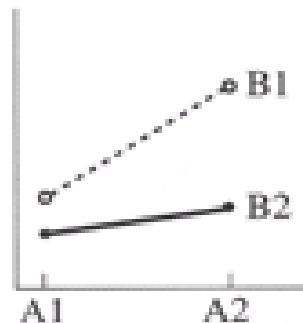
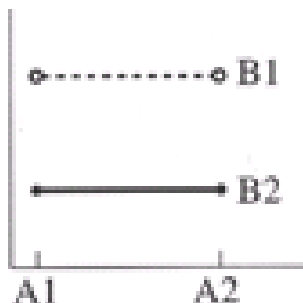
- Suppose are looking at two capsule types (C or V) & two digestive fluids (Gastric or Duodenal)
- Randomly assign 5 capsules of each type to each of type of digestive juice and dissolve time is observed and given below.

		Capsule Type	
		C	V
Type of Digestive Juice	Gastric	39.5	47.4
		45.7	43.5
		49.8	39.8
		50.2	36.1
		63.8	41.2
	Duodenal	33.5	44
		36.7	41.2
		42	47.3
		38.1	45.3
		31.2	42.7

Analysis of Variance – Two Way

Question of Interest:

- What effect does capsule type, fluid type or their interaction have on the time to dissolve? Which effect is most important?
- An interaction occurs when the effect of one factor depends on the level of another factor. Here we need to find out whether the effect of capsule depends on the type of digestive juice used to dissolve it or not?
- Interaction can be three type
 - **No interaction** – Effects are additive
 - **Synergistic interaction** – Interaction effects are added to main effects.
 - **Antagonistic interaction** - Interaction effects are subtracted from main effects.



Analysis of Variance – Two Way

Example of Interaction:

Refer to the capsule data:

```

> fit=aov(Time~Juice+Capsule+Juice*Capsule)
> summary(fit)

```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Juice	1	151.2	151.2	5.023	0.03954	*
Capsule	1	0.2	0.2	0.007	0.93605	
Juice:Capsule	1	320.0	320.0	10.628	0.00492	**
Residuals	16	481.8	30.1			

--- signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

- Interaction effect is highly significant.
- If interaction effect is significant, then choose the best combination from the Interaction effect plot (IEP) and not from Main Effect Plot (MEP).
- Residual check should be carried out before taking a decision on ANOVA

Analysis of Variance – Two Way

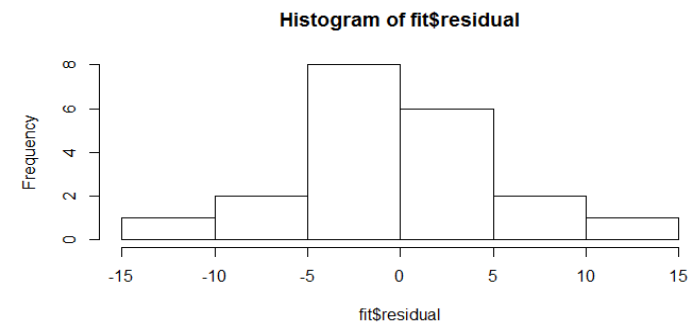
Checking Assumptions:

- Residual should follow Normal distribution
- Should not have any pattern against fitted value

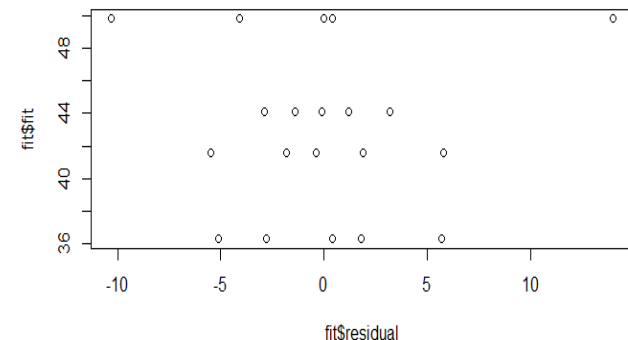
```
> hist(fit$residual)
➤ shapiro.test(fit$residual)
```

Shapiro-wilk normality test data:

```
fit$residual w = 0.94282, p-value = 0.2709
```



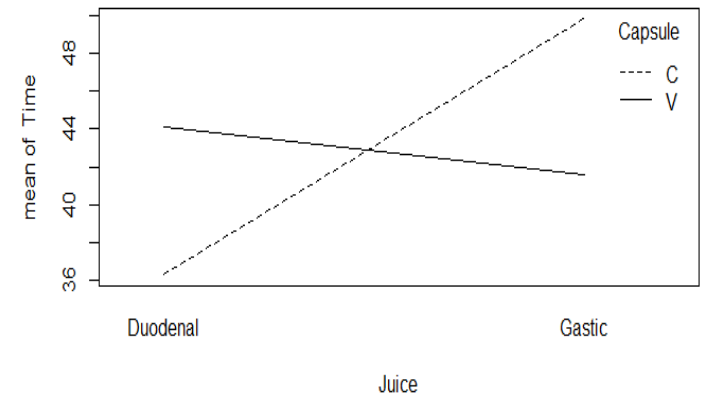
```
plot(fit$residual, fit$fit)
```



Analysis of Variance – Two Way

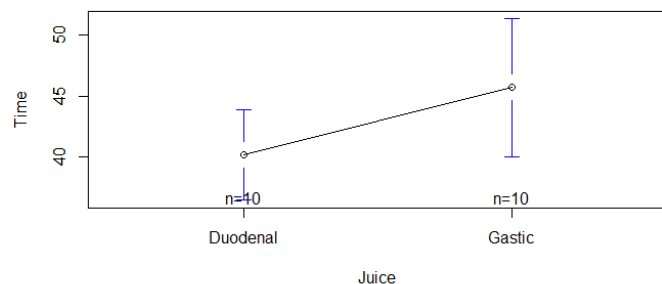
Interaction Effect Plot:

```
> interaction.plot(Juice,Capsule,Time,fun=mean)
```

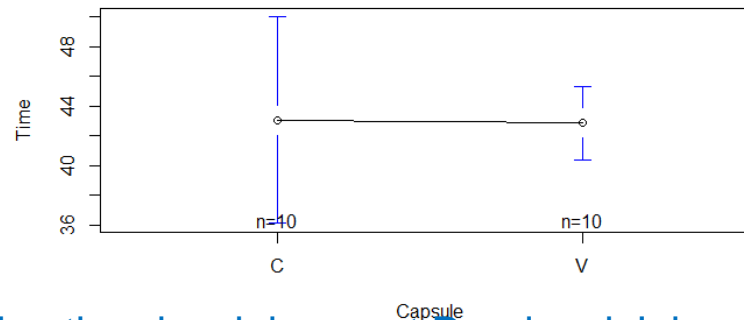


```
> library(gplots)
```

```
> plotmeans(Time~Juice)
```



```
> plotmeans(Time~Capsule)
```



As interaction Effect is significant, dissolving time is minimum at Duodenal Juice and C-type capsule combinations.

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?

1. 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)
```

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

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

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

Gender	Usage			Total
	1	2	3	
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

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

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

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

5. 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

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

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

```
> ftable(mytable)
```

Example: An experiment conducted, in which fish are placed in a large tank for a period of time and some are eaten by large birds of prey. The fish are categorized by their level of parasitic infection, either uninfected, lightly infected, or highly infected. It is to the parasites' advantage to be in a fish that is eaten, as this provides an opportunity to infect the bird in the parasites' next stage of life. The observed proportions of fish eaten are quite different among the categories.

Input Variable	Output Variable	
	Eaten	Not Eaten
Uninfected	1	49
Lightly infected	10	35
Highly Infected	37	9

The proportions of eaten fish are, respectively, $1/50 = 0.02$, $10/45 = 0.222$, and $37/46 = 0.804$.

- In the setting of the experiment, we observe a difference between the proportions of eaten fish in the lightly and highly infected fish. A point estimate of this difference is

$$37/46 - 10/45 = 0.804 - 0.222 = 0.582$$

- How can we quantify uncertainty in this estimate?

- A confidence interval for a difference in proportions $p_1 - p_2$ is based on the sampling distribution of the difference in sample proportions. If the two samples are independent,

$$E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$$

$$\text{Var}(\hat{p}_1 - \hat{p}_2) = \frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}$$

- If both samples are large enough (depending on how close the proportions are to 0 or 1), this sampling distribution is approximately normal.

95% Confidence Interval for $p_1 - p_2$

A 95% confidence interval for $p_1 - p_2$ is

$$\hat{p}_1 - \hat{p}_2 - 1.96\text{SE}(\hat{p}_1 - \hat{p}_2) < p_1 - p_2 < \hat{p}_1 - \hat{p}_2 + 1.96\text{SE}(\hat{p}_1 - \hat{p}_2)$$

where $\hat{p}_i = x_i/n_i$ for $i = 1, 2$ and

$$\text{SE}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

- For the infected fish case study, a confidence interval for the difference in probabilities of being eaten between highly and lightly infected fish is
 $0.415 < p(\text{high}) - p(\text{light}) < 0.749$

- Odds ratios are an alternative way to think about probabilities.
- Definition: The odds in favour of an event with probability p are $p/(1-p)$.
- The odds ratio in favor of an event between two groups is the odds in favor for the first group divided by the odds in favor for the second group.

$$\text{Odds Ratio} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}$$

- Odds ratios are estimated by plugging in sample proportions.
- The estimated odds for being eaten in the highly infected group is $37/9 = 4.111$.
- The estimated odds for being eaten in the lightly infected group is $10/35 = 0.286$.
- The estimated odds ratio is 14.389

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 depends 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					Row Sum
	VD	D	N	S	VS	
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

$$= (140 \times 61) / 400 = 21.4$$

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

Experience (Months)	CSat Score					Row Sum
	VD	D	N	S	VS	
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 $(\text{observed} - \text{expected})^2 / \text{expected}$ for each cell

Table of $(\text{observed} - \text{expected})^2 / \text{expected}$

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 $((\text{observed} - \text{expected})^2 / \text{expected})$

Chi Square calculated = $38.45 + 6.32 + \dots + 34.5$

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 = $\text{chidist}(\text{chi Sq}, \text{df})$

= $\text{chidist}(161.98, 8)$

= 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

In this example - **Phi Coefficient** = $\sqrt{161.98 / 400} = 0.64$

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

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

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

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			Total
	1	2	3	
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)
```

	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

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 customer's 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

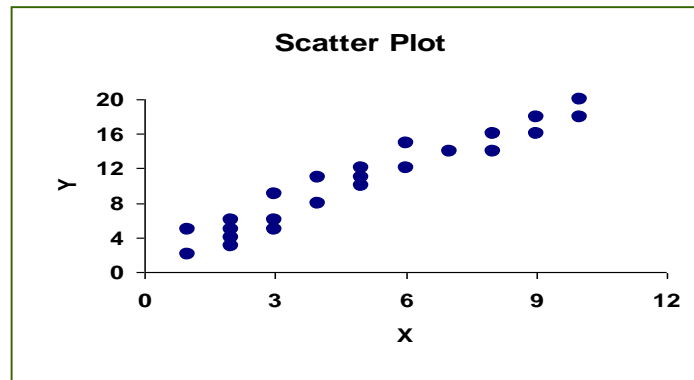
Explore the relationship between the output characteristic and input or process variable.

Output variable : Y : Dependent variable

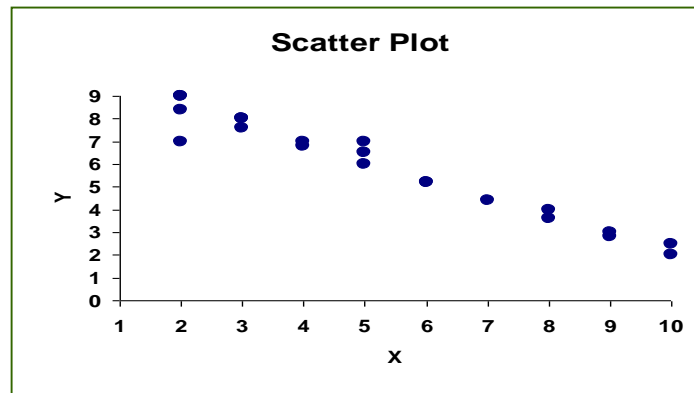
Input / Process variable : X : Independent variable

CORRELATION & REGRESSION

Positive Correlation: Y increases as X increases & vice versa

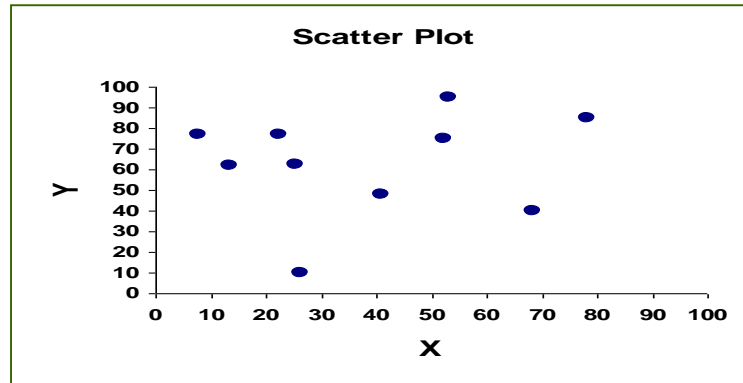


Negative Correlation: Y decreases as X increases & vice versa

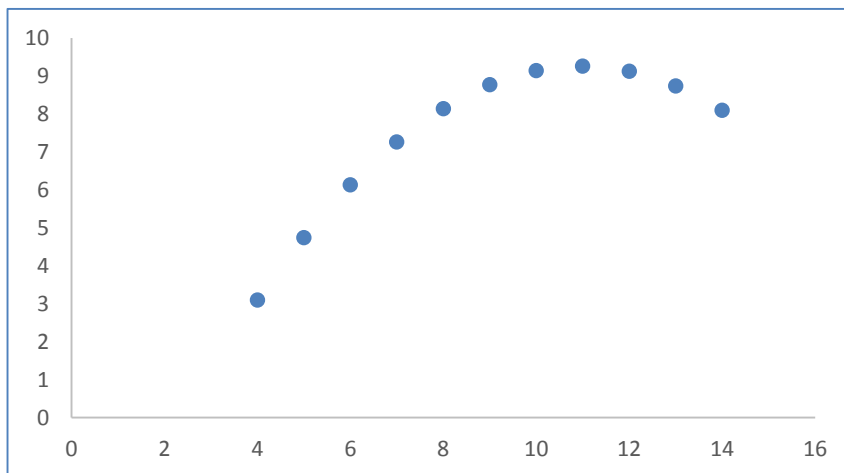


CORRELATION & REGRESSION

No Correlation: Random Distribution of points

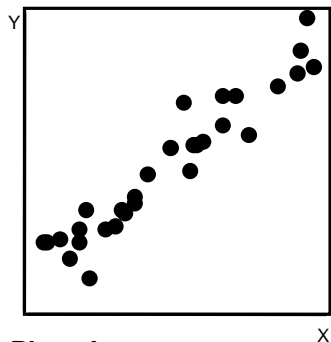


Non Linear Correlation: Curvature form of points

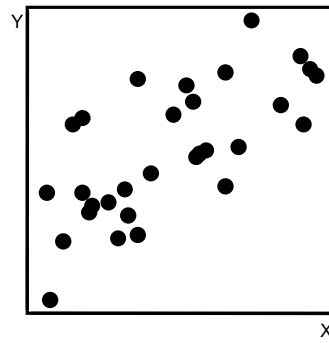


CORRELATION & REGRESSION

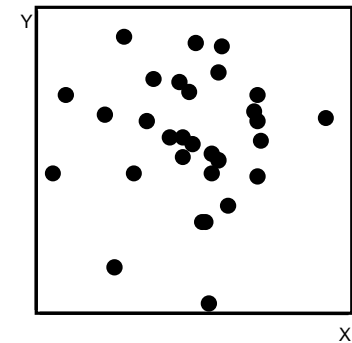
Is there any correlation ?



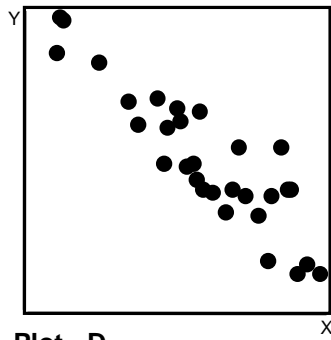
Plot - A



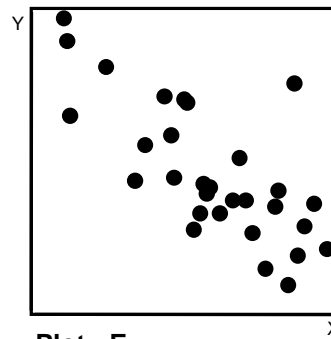
Plot - B



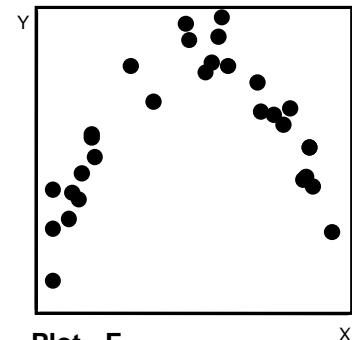
Plot - C



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

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

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

CORRELATION & REGRESSION

Coefficient of Correlation Computation :

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

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

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

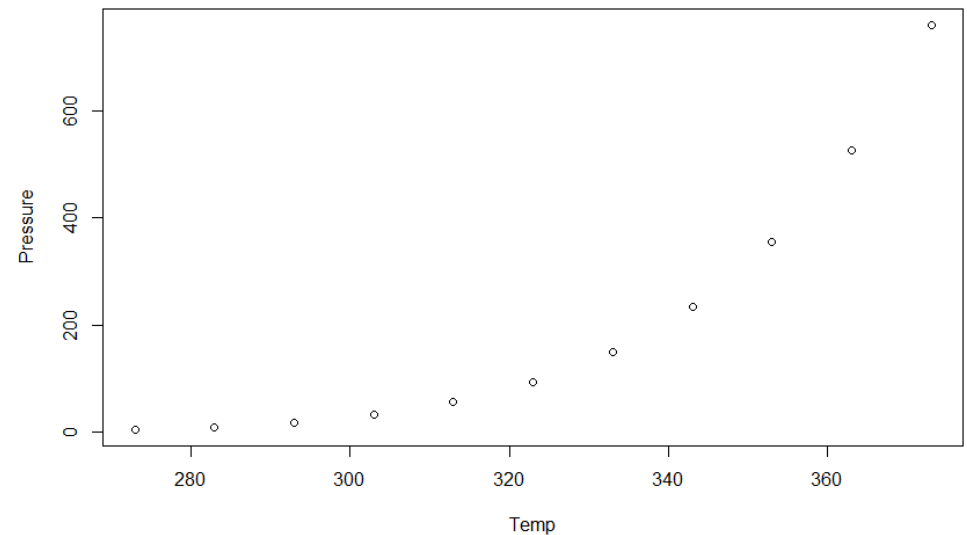
```
> mydata = Correlation  
> Temp = mydata$Temperature  
> Pressure = mydata$Vapor.Pressure
```

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) = $5 + 3 \times \text{Time} - 2 \times \text{Temperature}$

CORRELATION & REGRESSION

Simple Linear Regression Illustration

Output variable is modeled in terms of only one variable

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

Regression Model

$$y = 1 + 3x$$

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

$$H_0: b = 0$$

$$H_1: b \neq 0$$

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

If $p \text{ value} < 0.05$, then H_0 is rejected & y can be modeled with x

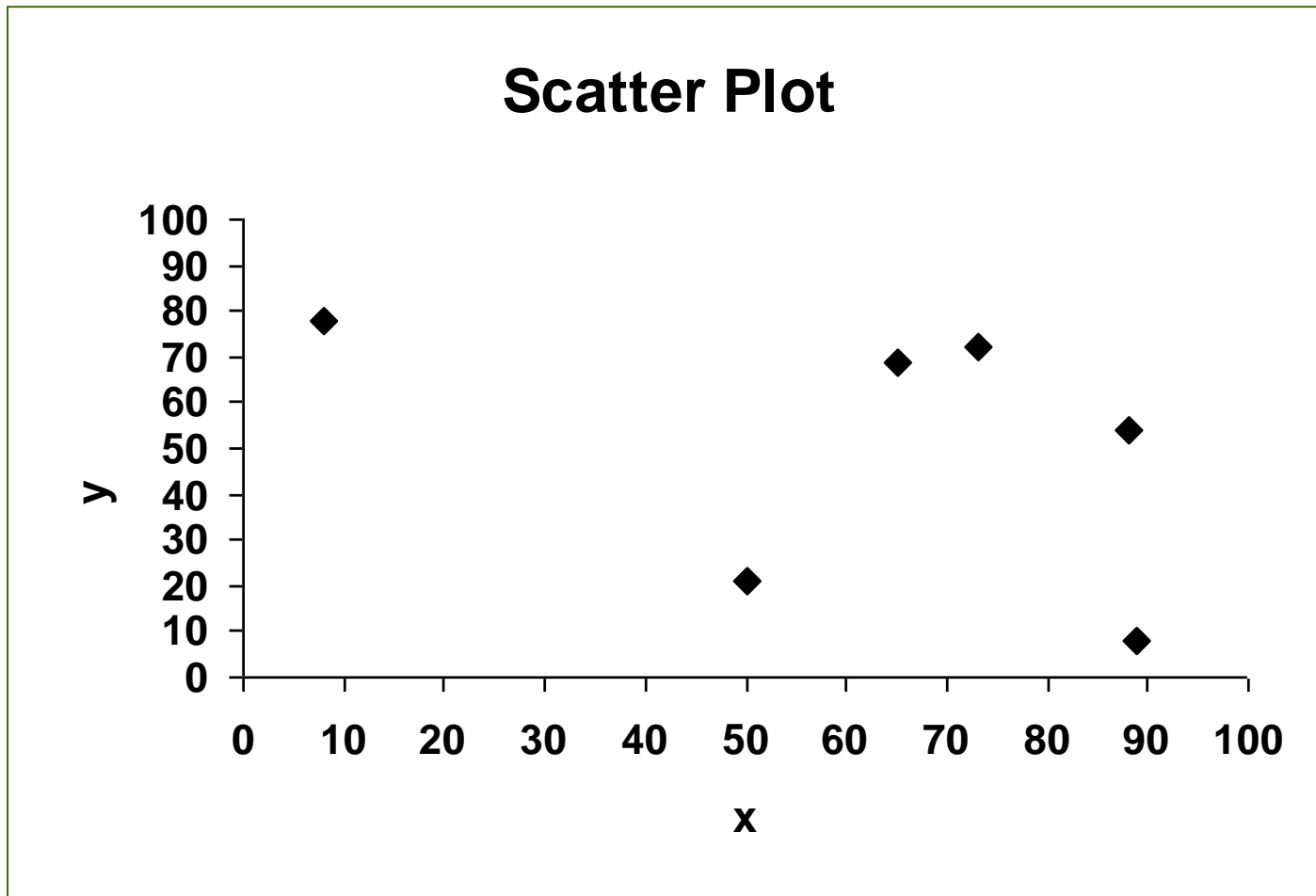
CORRELATION & REGRESSION

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

$$R^2 = SS_R / S_{yy} = b \cdot S_{xy} / S_{yy}$$

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

$$S_{yy} = \sum (y_{\text{actual}} - \text{Mean } y)^2$$

R^2 : amount variation in y explained by x

Range of R^2 : 0 to 1

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

CORRELATION & REGRESSION

Coefficient of Regression: Testing the significance of Regression

Regression ANOVA

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

If $p \text{ value} < 0.05$, 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.

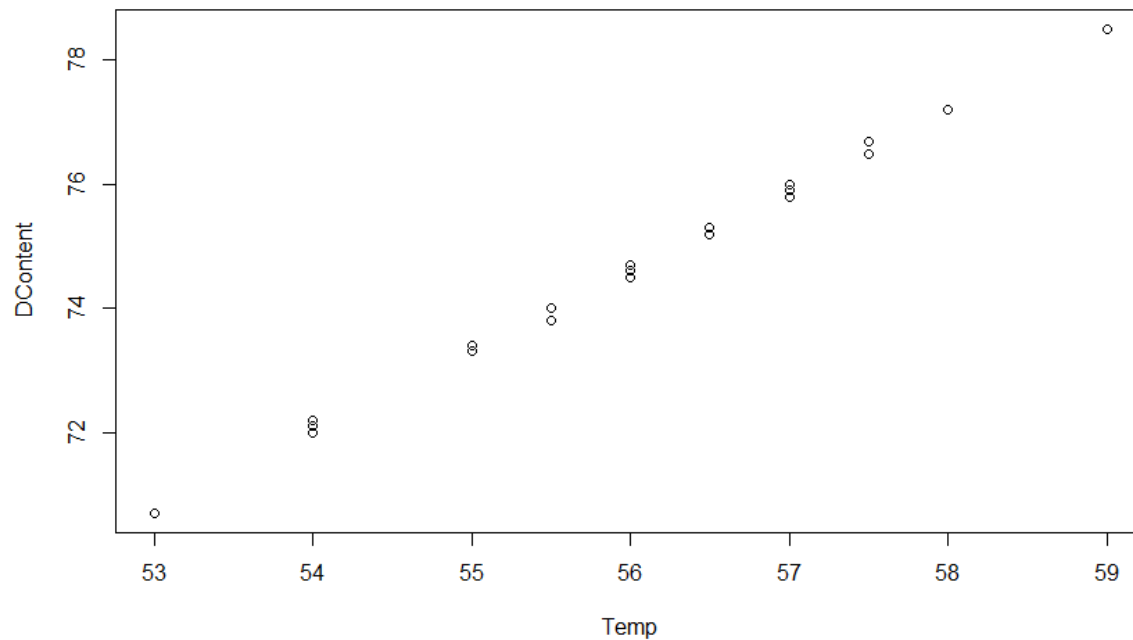
1. Reading the data and variables

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

CORRELATION & REGRESSION

2. Constructing Scatter Plot

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



CORRELATION & REGRESSION

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)

CORRELATION & REGRESSION

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

CORRELATION & REGRESSION

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

CORRELATION & REGRESSION

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")
```

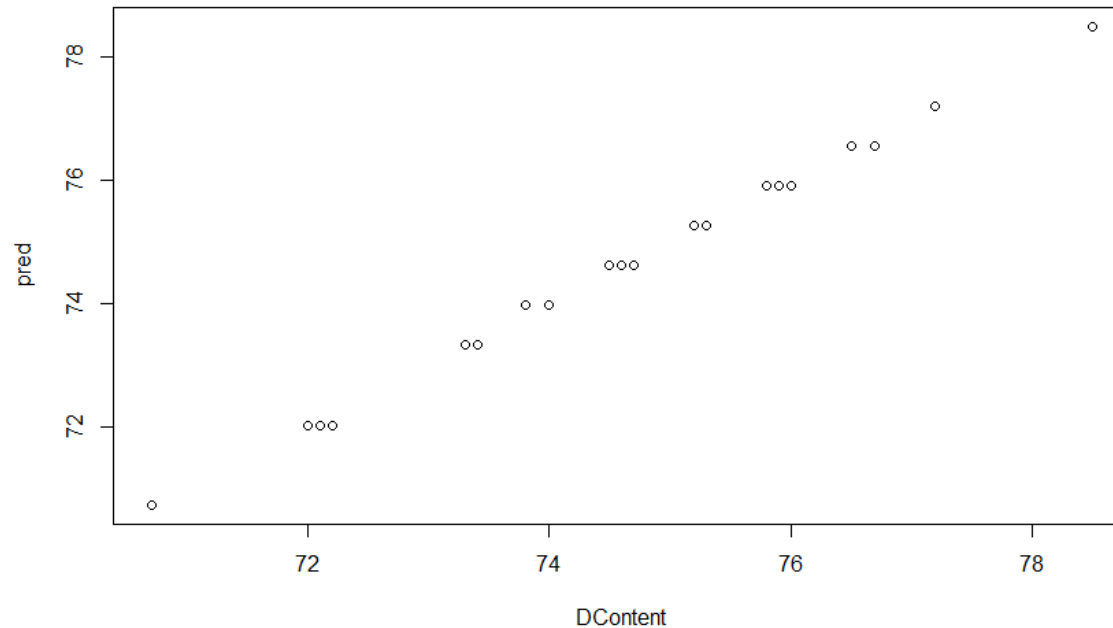
SL No.	Fitted	Residuals	SL No.	Fitted	Residuals
1	73.32259	-0.02259	22	74.61602	-0.01602
2	74.61602	-0.01602	23	75.26274	-0.06274
3	73.96931	0.030693	24	73.96931	0.030693
4	78.49632	0.00368	25	75.90946	-0.00946
5	74.61602	-0.01602	26	75.26274	0.03726
6	73.96931	0.030693	27	73.96931	0.030693
7	75.26274	-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

CORRELATION & REGRESSION

5: Residual Analysis

Scatter Plot: Actual Vs Predicted (fit)

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



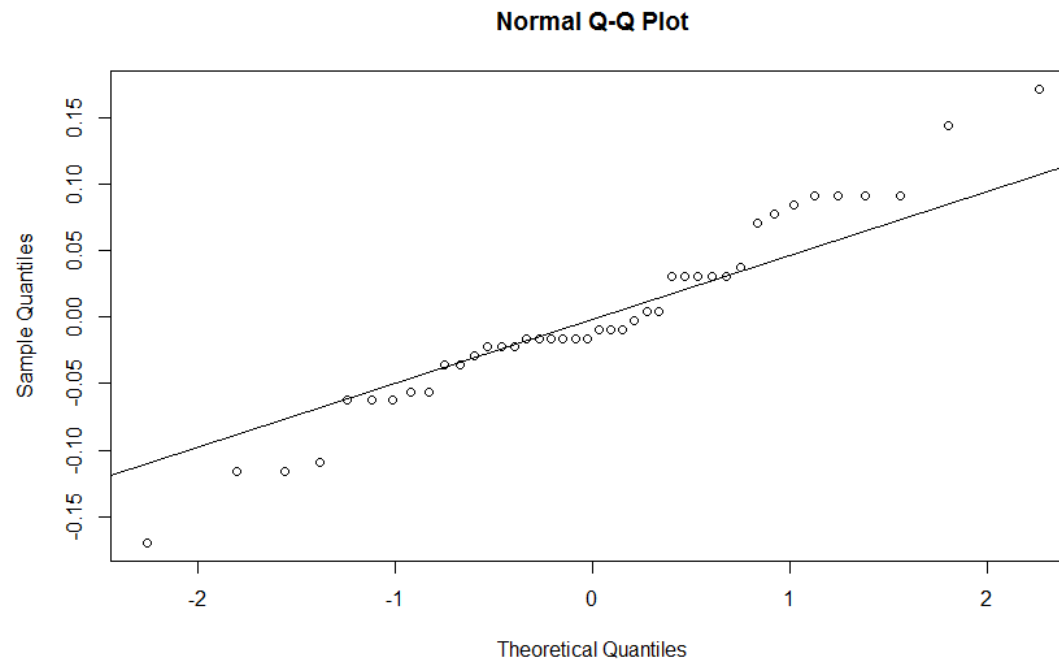
CORRELATION & REGRESSION

5: Residual Analysis

Normality Check on residuals

```
> qqnorm(Res)
```

```
> qqline(Res)
```



Residuals should be normally distributed or bell shaped

CORRELATION & REGRESSION

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

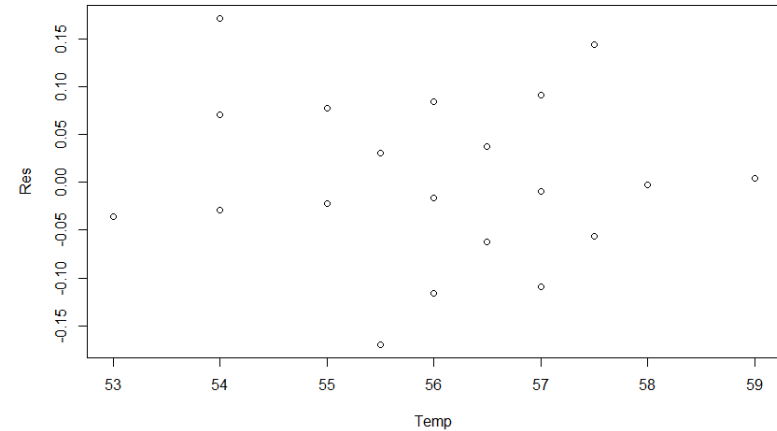
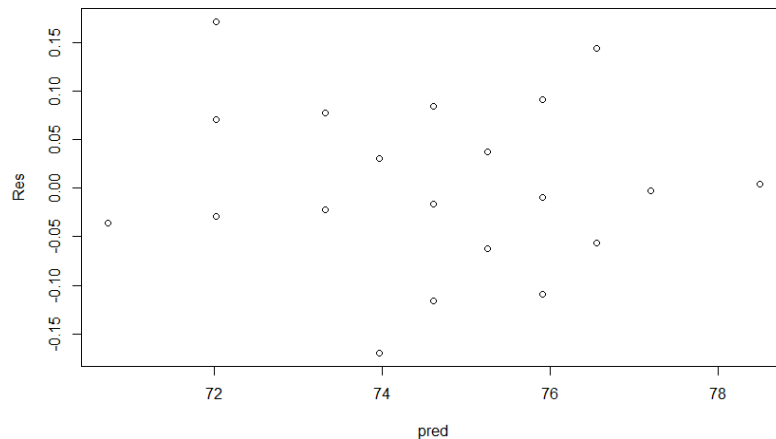
Plot the residuals against fitted value. The points in the graph should be scattered randomly and should not show any trend or pattern. The residuals should not depend in anyway on the fitted value.

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.

CORRELATION & REGRESSION

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

REGRESSION ANALYSIS

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 $MSE_1 = (y_{1\text{actual}} - y_{1\text{pred}})^2$

- Repeat the process by taking (x_2, y_2) as test data and the remaining $n - 1$ observations as training data
- Compute MSE_2
- Repeating the procedure n times produces n squared errors $MSE_1, MSE_2, \dots, MSE_n$
- LOOCV estimate of the test MSE is the average of these n test error estimates

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

REGRESSION ANALYSIS

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:

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

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, $j = 1, 2, \dots, k$)

CORRELATION & REGRESSION

Exercise : The effect of temperature and reaction time affects the % yield. The data collected is 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 x s & y should be high

Correlation between x s 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

CORRELATION & REGRESSION

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

CORRELATION & REGRESSION

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 $p \text{ value} < 0.05$

CORRELATION & REGRESSION

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

CORRELATION & REGRESSION

Step 3: Residual Analysis:

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

Step 4: Outlier test

Observations with Bonferonni p – value < 0.05 are potential outliers

```
> library(car)
```

```
> outlierTest(mymodel)
```

Observation	Studentized Residual	Bonferonni p value
11	1.781515	NA

REGRESSION ANALYSIS

5: 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 Cl₂ 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**

CORRELATION & REGRESSION

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

REGRESSION ANALYSIS

Tackling Multicollinearity:

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

REGRESSION ANALYSIS

Tackling Multicollinearity:

Method 1: Removing highly correlated variable – Stepwise Regression

Approach

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

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

REGRESSION ANALYSIS

Tackling Multicollinearity:

Method 1: Removing highly correlated variable – Stepwise Regression

Akaike information criterion (AIC)

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

n: number of observations

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

d: number of x variables included in the model

RSS: Residual sum of squares

REGRESSION ANALYSIS

Tackling Multicollinearity:

Method 1: Removing highly correlated variable – Stepwise Regression

R code

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

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

REGRESSION ANALYSIS

Tackling Multi collinearity:

Method 1: Stepwise Regression

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

$$\% \text{ Conversion} = 0.13396 * \text{Temperature} + 0.35106 * \text{Kappa \#} - 130.68986$$

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

REGRESSION ANALYSIS

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

REGRESSION ANALYSIS

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

CORRELATION & REGRESSION

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?

CORRELATION & REGRESSION

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

CORRELATION & REGRESSION

Regression with dummy variables

Read the file and variables

```
> mydata = read.csv("Travel_dummy_Reg.csv")
```

```
> 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)
```

CORRELATION & REGRESSION

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 $y = x\beta + \varepsilon$ is general model

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

Examples:

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

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

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

In general

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

where $f(x)$ can be $1/x$, \sqrt{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 thrust 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

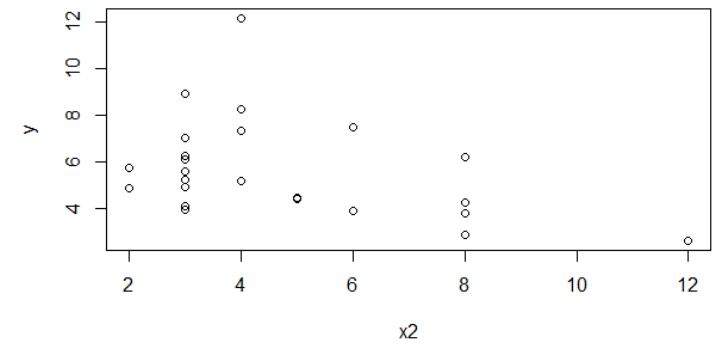
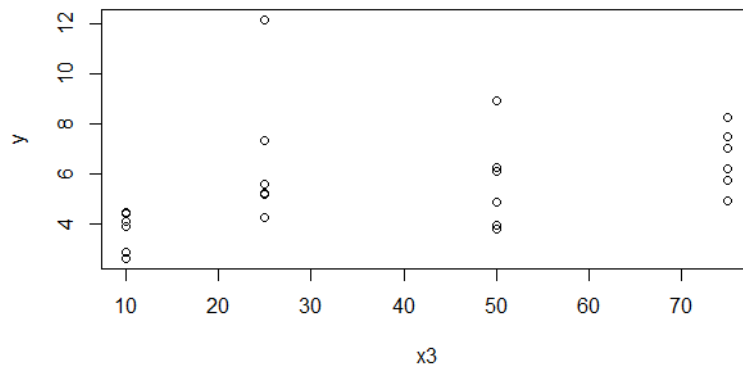
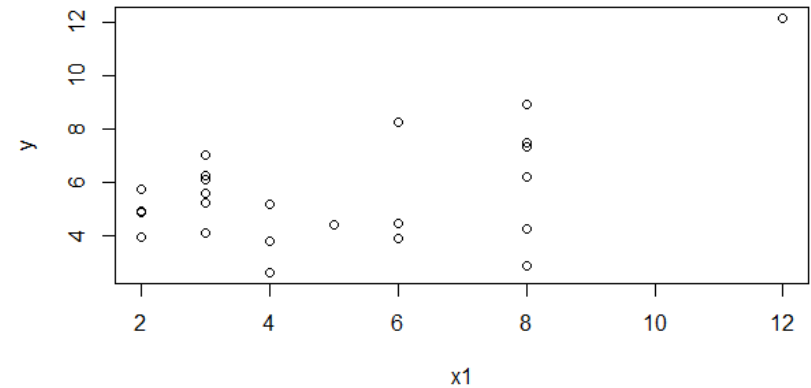
MODELING NONLINEAR RELATIONS

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

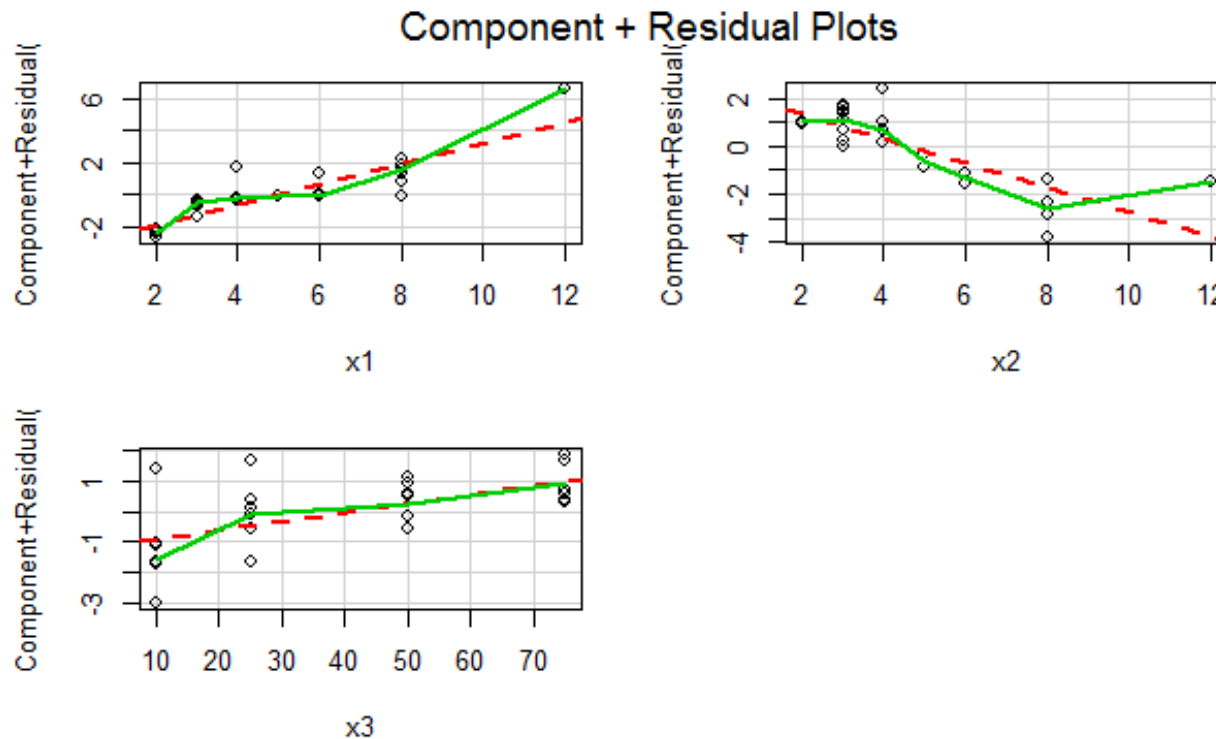
R ²	0.786
Adjusted R ²	0.7563

MODELING NONLINEAR RELATIONS

Develop the model

```
> library(car)
```

```
> crPlots(mymodel))
```



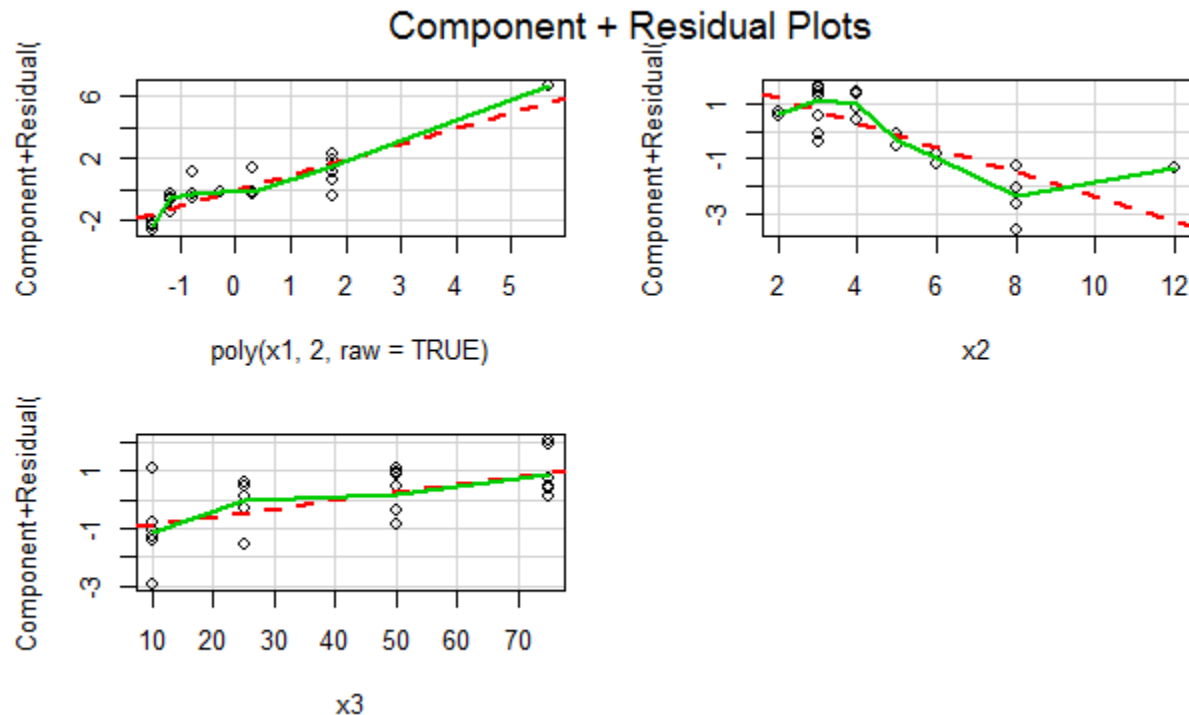
Since the best fit line different from residual line, it is possible improve the model by adding higher order terms

MODELING NONLINEAR RELATIONS

Develop the model

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

```
> crPlots(mymodel)
```

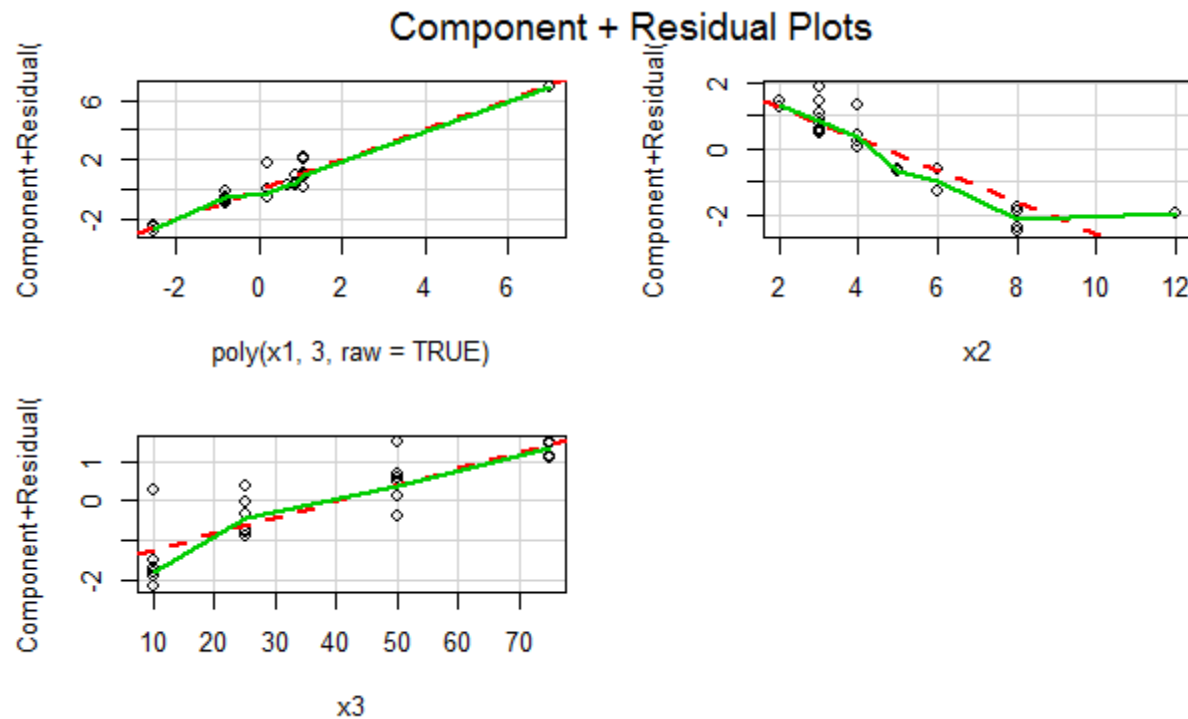


Since the best fit line different from residual line, it is possible improve the model by adding higher order terms

MODELING NONLINEAR RELATIONS

Develop the model

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

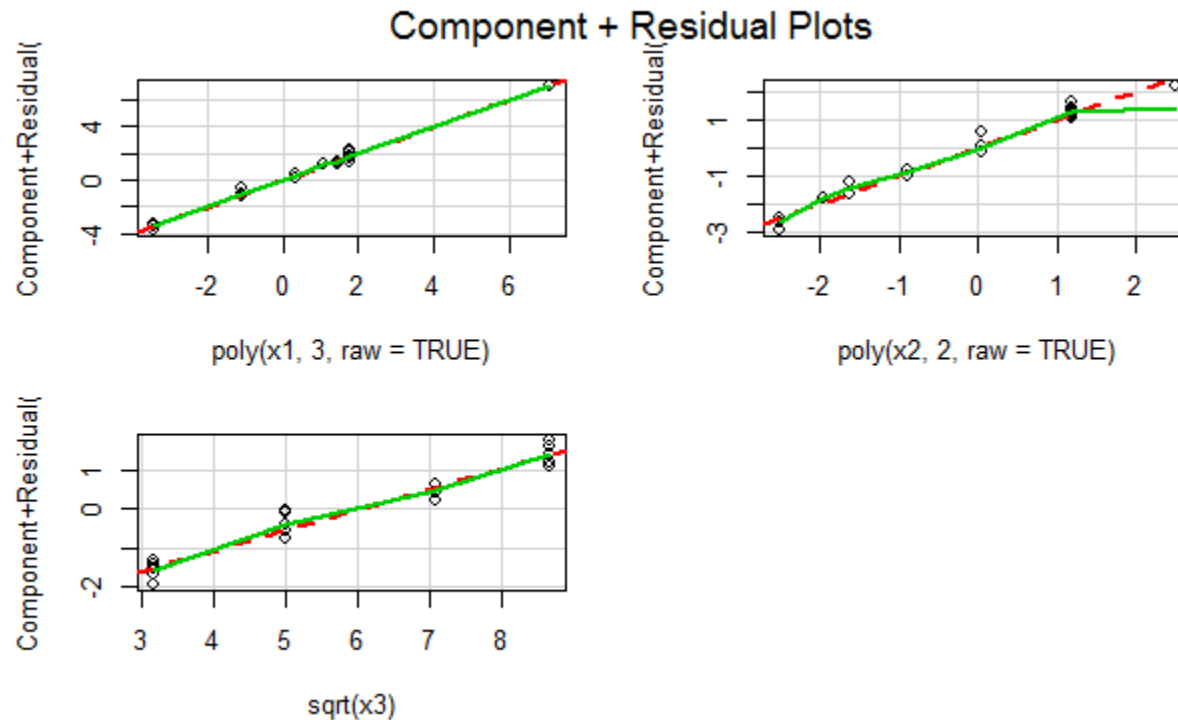


Since the best fit line is more or less overlapping residual line, hence adding square and cube terms of $x1$ will improve the model. Similarly add additional terms or functions of $x2$ and $x3$ to improve the model

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

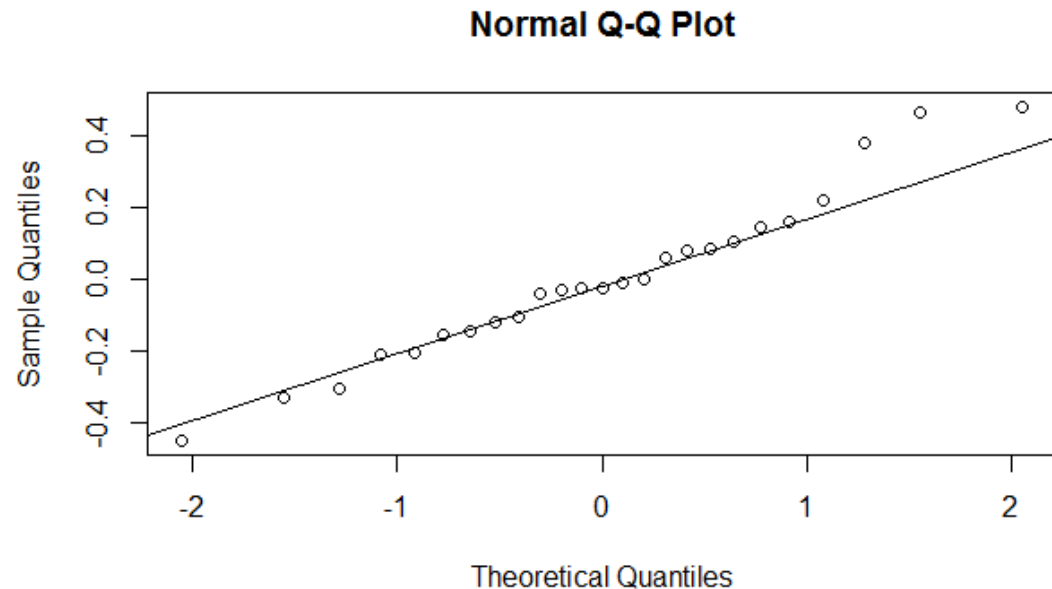
Develop the model: **Final Model**

```
> 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

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

p : probability of success

x_i 's : independent variables

a, b_1, b_2, \dots : 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)
```

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

4. Converting response variable to discrete

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


BINARY LOGISTIC REGRESSION

5a. Checking relation between Xs and Y

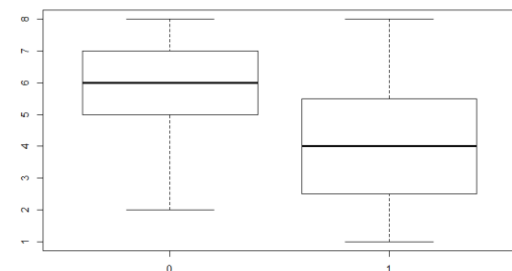
- > aggregate(income ~visit, FUN = mean)
- > aggregate(attitude ~visit, FUN = mean)
- > aggregate(importance ~visit, FUN = mean)
- > aggregate(size ~visit, FUN = mean)
- > aggregate(age ~visit, FUN = mean)

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

5b. Checking relation between Xs and Y – box plot

- > boxplot(income ~ 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 $p \text{ value} < 0.05$ for Income, Importance_Vacation & Size, redo the modelling with important factors only

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

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

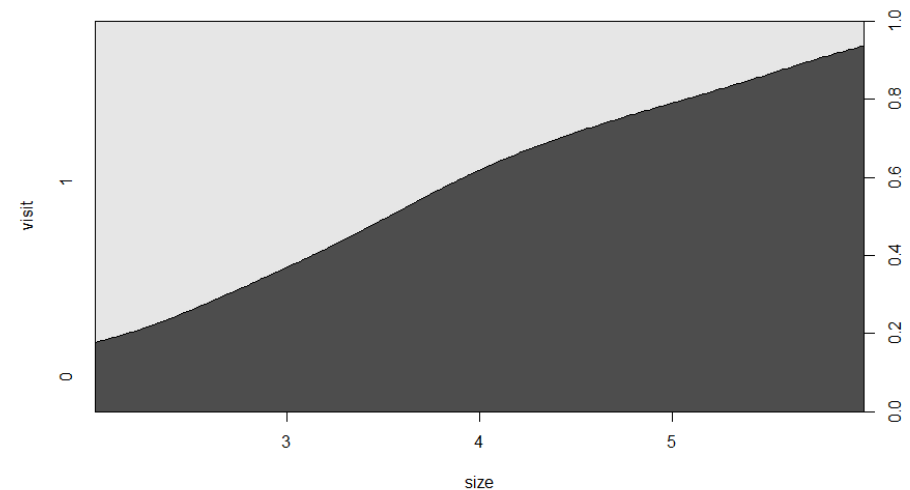
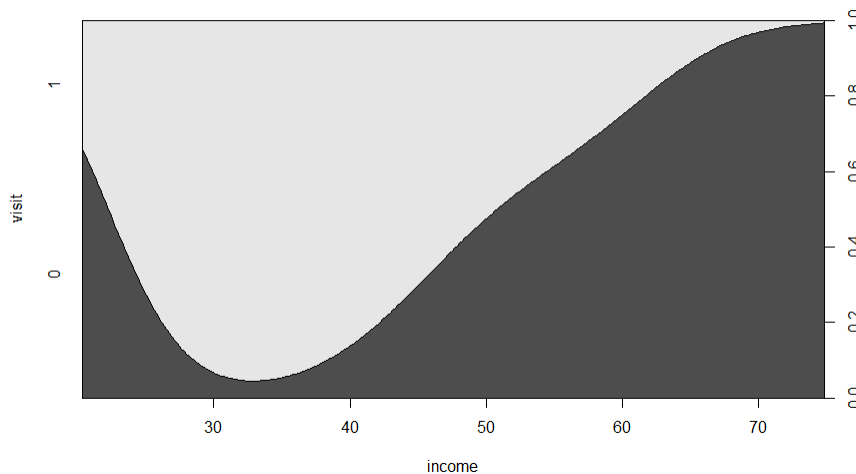
BINARY LOGISTIC REGRESSION

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)
```



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

BINARY LOGISTIC REGRESSION

10. Model Evaluation

```
> mytable = table(vvisit, 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 $\geq 80\%$ is good

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 numeric

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:

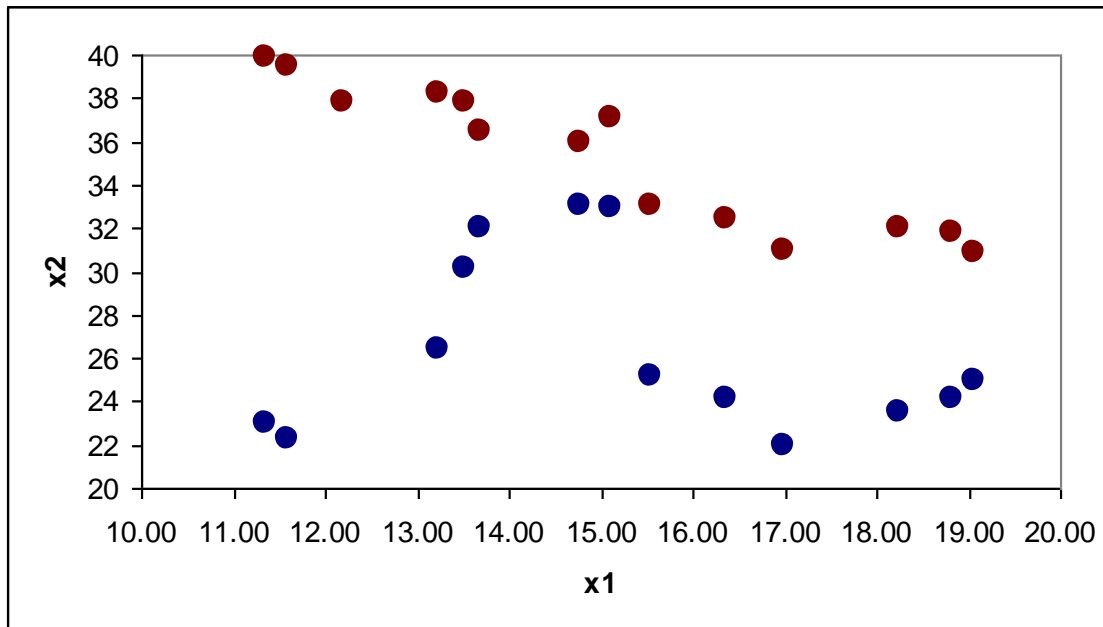
Attribute 1	x1
Attribute 2	x2
Label : y	Y1 (Red) , y2 (Blue)

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:

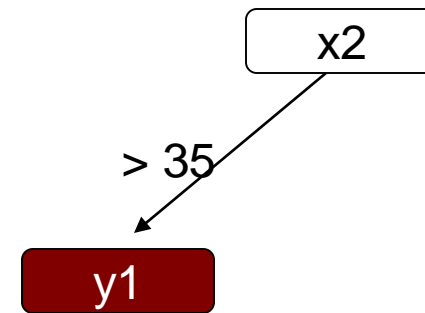
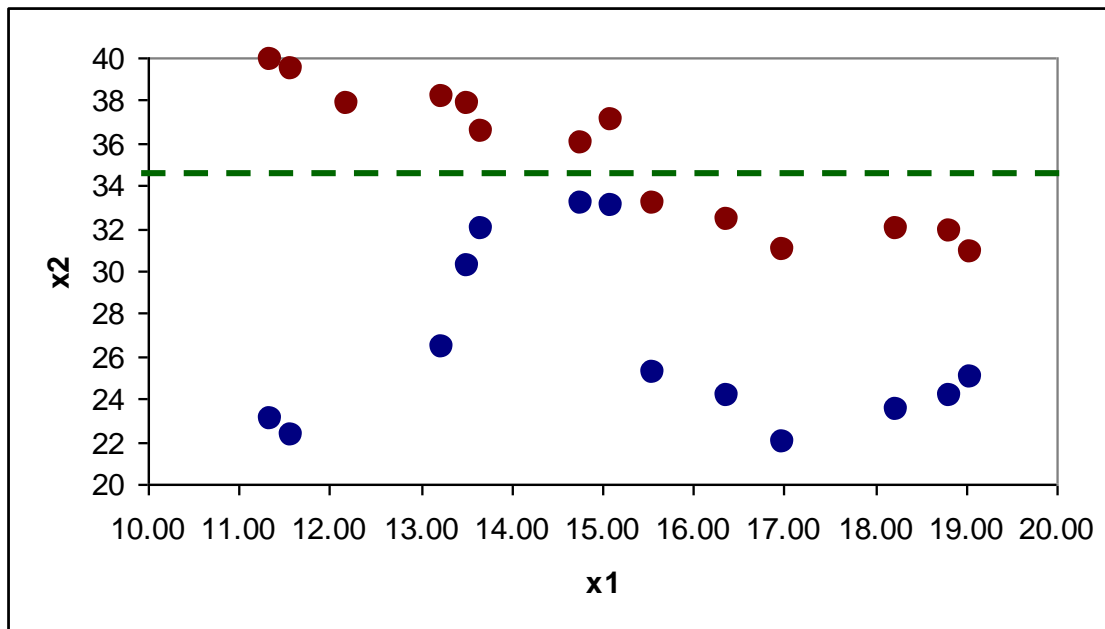
Attribute 1	x1
Attribute 2	x2
Label : y	Y1 (Red) , y2 (Blue)



CLASSIFICATION AND REGRESSION TREE

Example:

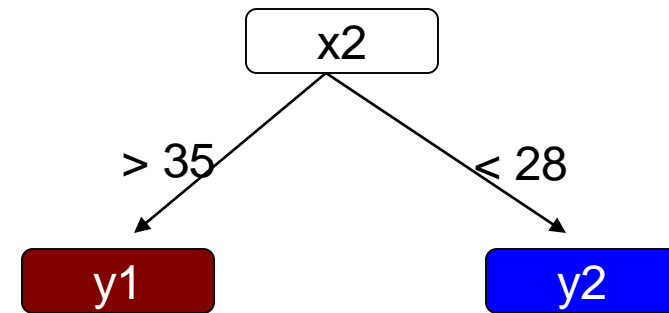
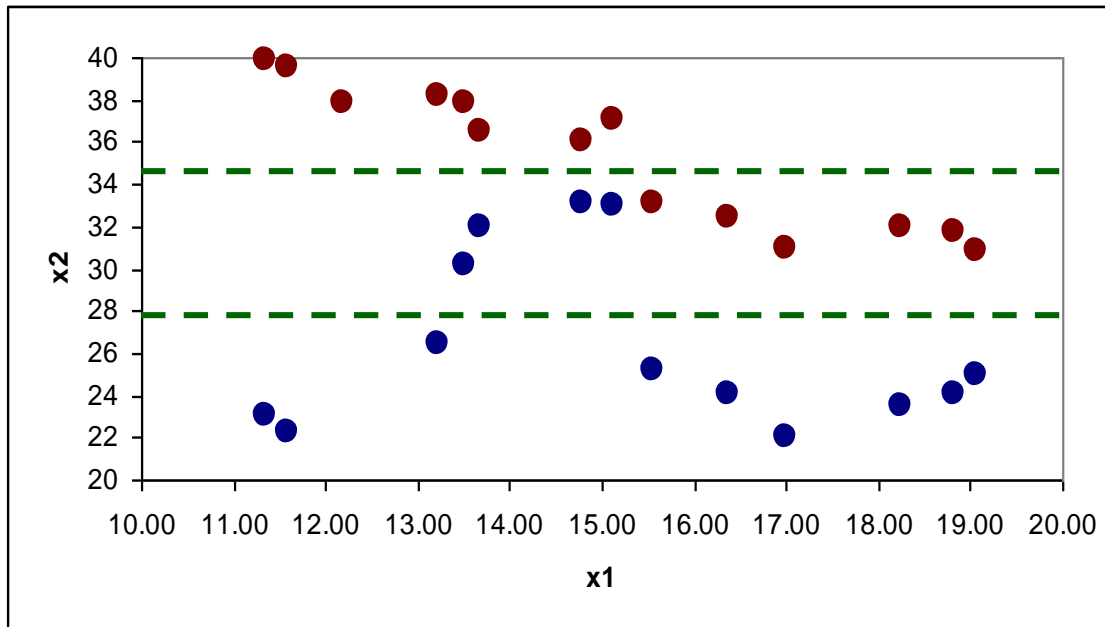
Attribute 1	x1
Attribute 2	x2
Label : y	y1 (Red) , y2 (Blue)



CLASSIFICATION AND REGRESSION TREE

Example:

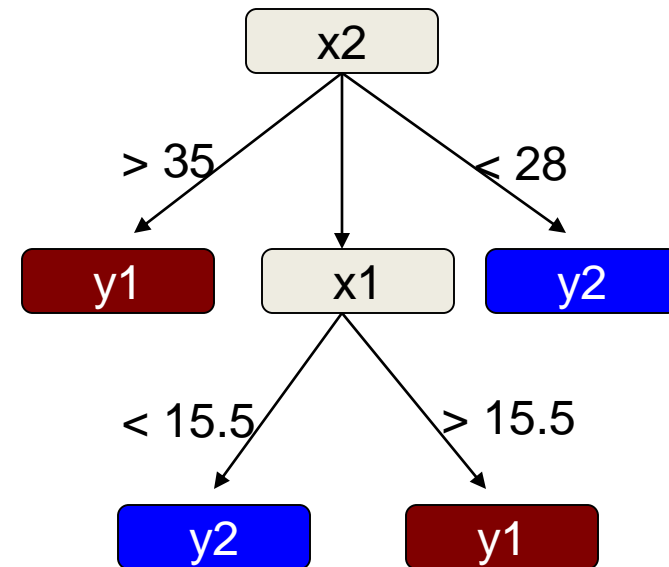
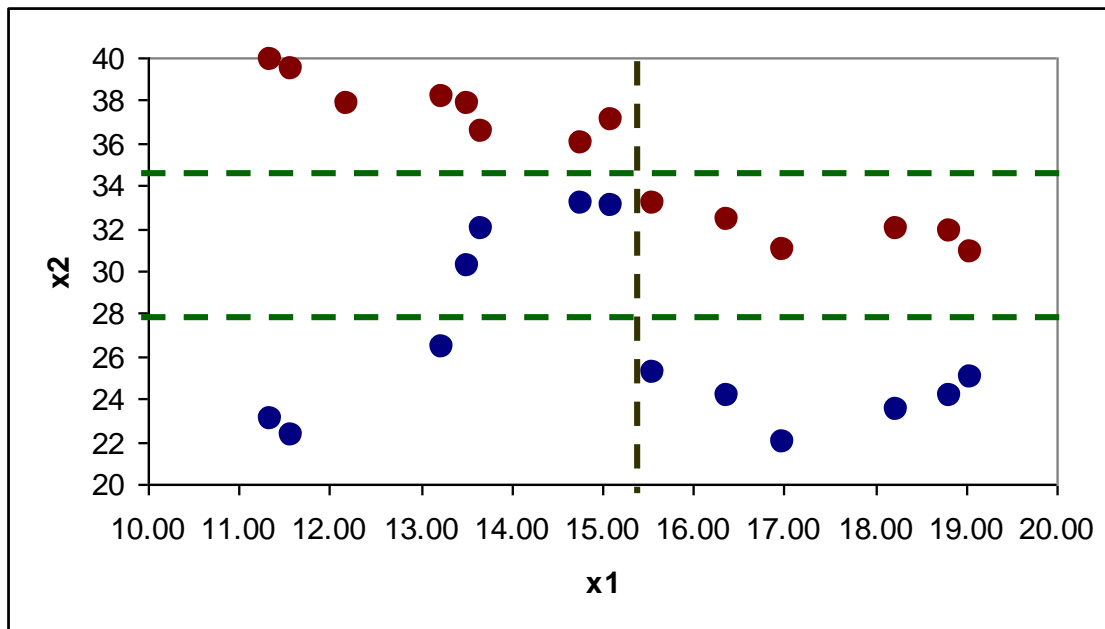
Attribute 1	x1
Attribute 2	x2
Label : y	y1 (Red) , y2 (Blue)



CLASSIFICATION AND REGRESSION TREE

Example:

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



CLASSIFICATION AND REGRESSION TREE

Example: Rules

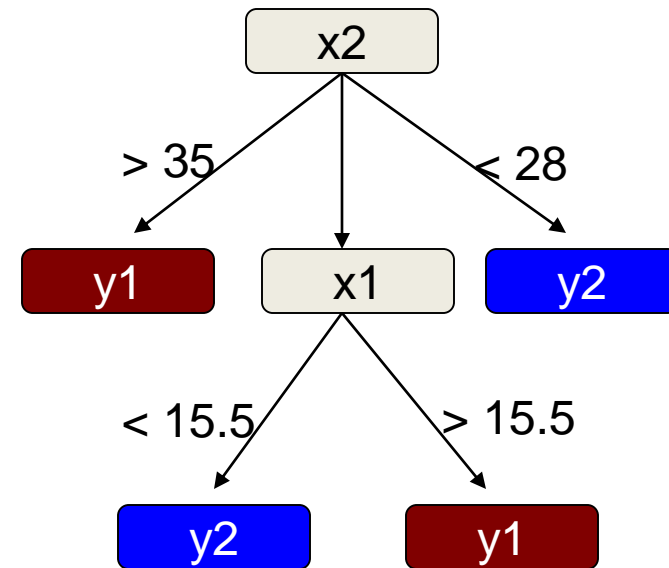
Attribute 1	x1
Attribute 2	x2
Label : y	y1 (Red) , y2 (Blue)

If $x_2 > 35$ then $y = y_1$

If $x_2 < 28$, then $y = y_2$

If $28 > x_2 > 35$ & $x_1 > 15.5$, then $y = y_1$

If $28 > x_2 > 35$ & $x_1 < 15.5$, then $y = y_2$



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 in mail_respond.csv? Can you develop a rule to identify the profile of customers who are likely to respond?

Number of variables = 4

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

Total Combination of Customer Profiles = $3 \times 3 \times 2 \times 2 = 36$

CLASSIFICATION AND REGRESSION TREE

Read file and variables

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

CLASSIFICATION AND REGRESSION TREE

Develop the model

```
> library(rpart)
```

```
> library(rpart.plot)
```

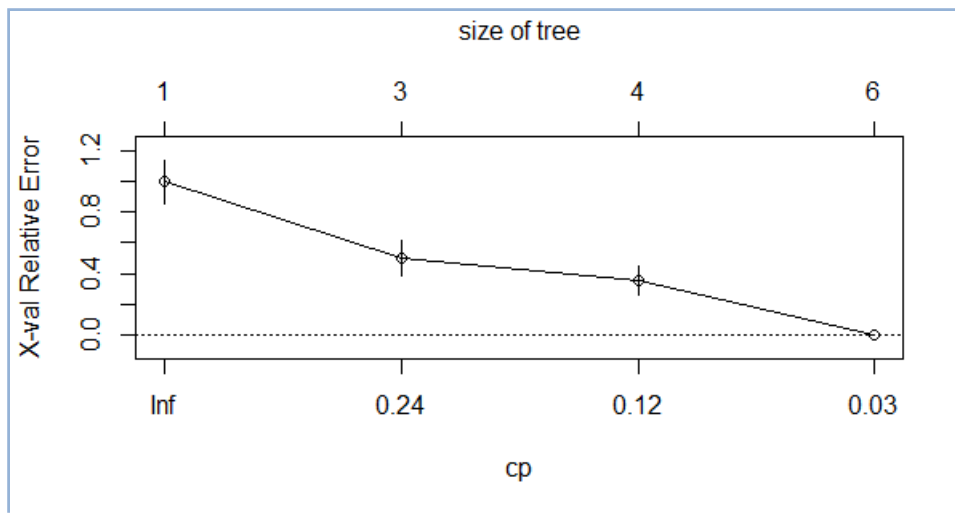
```
> mymodel = rpart(outcome ~ district + house + income + prev, method = 'class', control =  
rpart.control(minsplit = 10))
```

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

CLASSIFICATION AND REGRESSION TREE

Cross validation and identification of cp

```
> plotcp(mymodel)
```



Optimum cp = **0.03**

Corresponding to minimum cross validation relative error

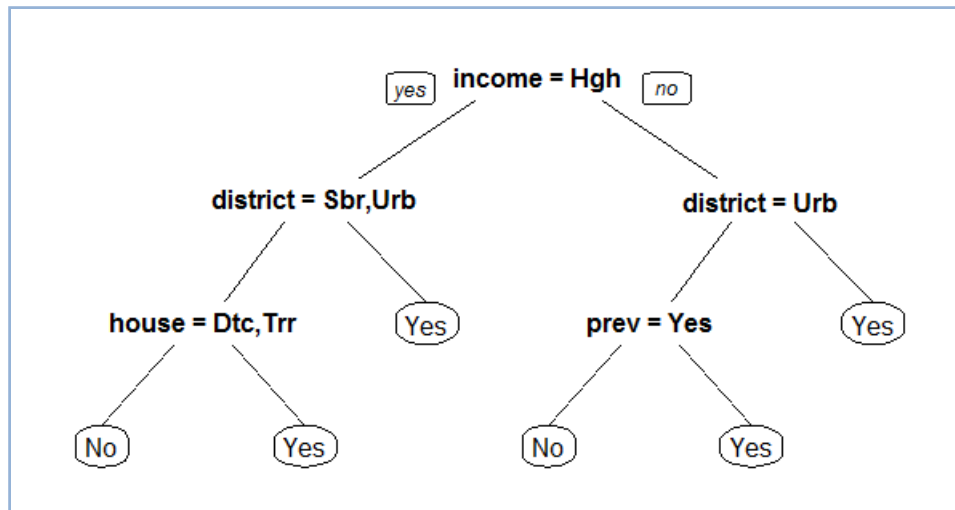
CLASSIFICATION AND REGRESSION TREE

Pruning the model

```
> mymodel = prune(mymodel, cp = 0.03)
```

Plot the model

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



CLASSIFICATION AND REGRESSION TREE

Print the model

```
> mymodel
```

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

CLASSIFICATION AND REGRESSION TREE

Model Accuracy measures

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

```
> mytable = table(outcome, pred)
```

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

Actual Vs predicted: %

Actual	Predicted	
	No	Yes
No	34	0
Yes	0	66

$$\text{Accuracy} = 34 + 66 = 100\%$$

CLASSIFICATION AND REGRESSION TREE

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

CLASSIFICATION AND REGRESSION TREE

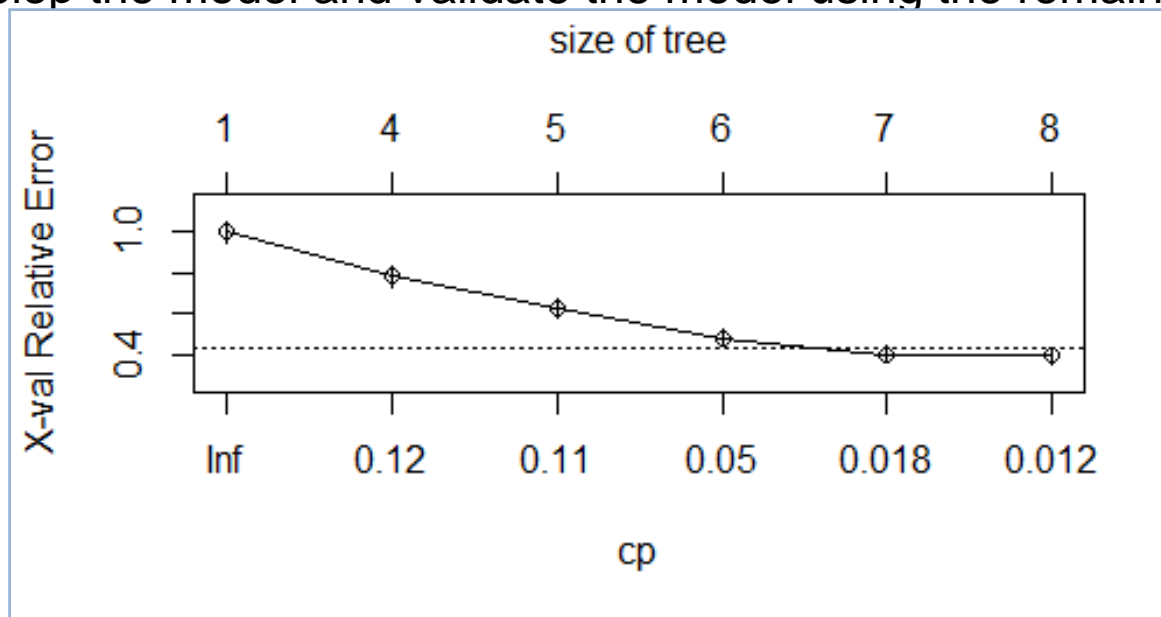
Exercise 1: Develop a tree based model for predicting whether the customer will take pep using the customer profile data given in bank-data.csv? Use 80% of data to develop the model and validate the model using the remaining 20% of data?

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

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

CLASSIFICATION AND REGRESSION TREE

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



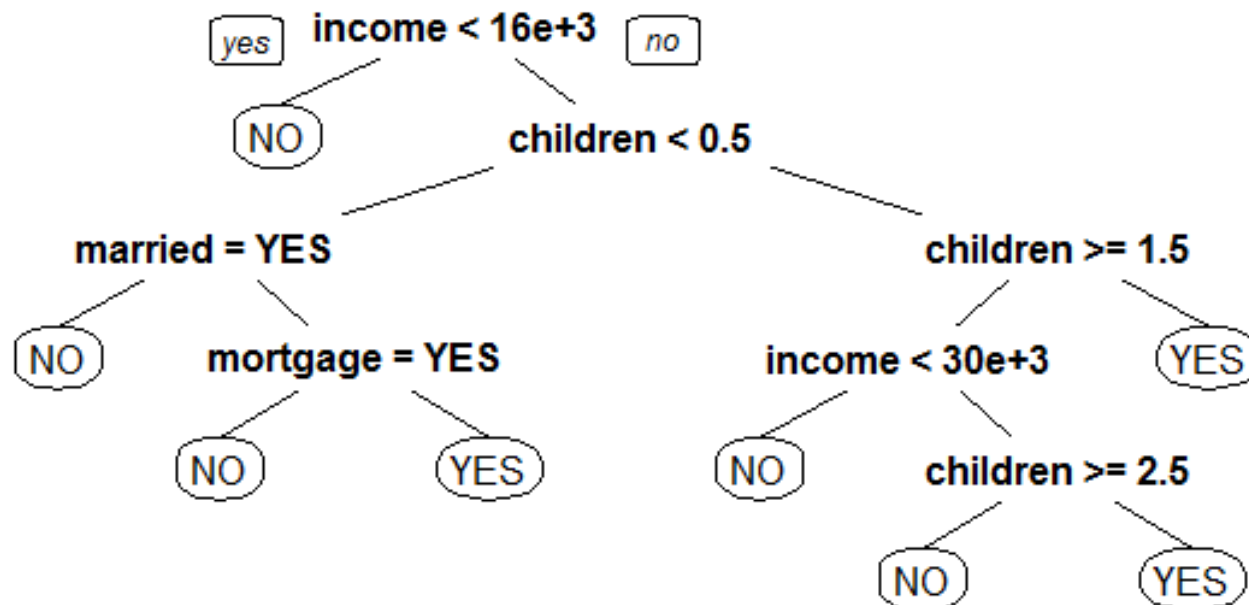
Optimum cp with minimum cross validation relative error = **0.012**

```
> mymodel = prune(mymodel, cp = 0.012)
```

CLASSIFICATION AND REGRESSION TREE

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

`rpart.plot(mymodel)`



CLASSIFICATION AND REGRESSION TREE

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

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

Actual Vs predicted: %

Actual	Predicted	
	No	Yes
No	52.89	1.65
Yes	13.64	31.82

Accuracy = 52.89 + 31.82 = 84.71 %

CLASSIFICATION AND REGRESSION TREE

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

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

Actual Vs predicted: %

Actual	Predicted	
	No	Yes
No	50.86	2.59
Yes	18.97	27.59

Accuracy = $50.86 + 27.59 = 78.45 \%$

CLASSIFICATION AND REGRESSION TREE

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

CLASSIFICATION AND REGRESSION TREE

Random Forest

Improves predictive accuracy

Generates large number of bootstrapped trees

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

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

Regression tree – average

Classification tree – majority vote

CLASSIFICATION AND REGRESSION TREE

Random Forest

Example

Develop a model to predict plant class using the data given in Iris.csv.\ using random forest method. Validate the model with Iris_test.csv data?

CLASSIFICATION AND REGRESSION TREE

Random Forest

Example

Develop a model to predict plant class using the data given in Iris.csv.\ using random forest method. Validate the model with Iris_test.csv data?

CLASSIFICATION AND REGRESSION TREE

Random Forest

Example

Develop a model to predict plant class using the data given in Iris.csv.\ using random forest method. Validate the model with Iris_test.csv data?

Read Iris data to mydata

```
> library(randomForest)
> mymodel = randomForest(Class ~ sepal.length + sepal.width + petal.length + petal.width, data
= mydata)
> mymodel
```

CLASSIFICATION AND REGRESSION TREE

Random Forest

Example

Develop a model to predict plant class using the data given in Iris.csv.\ using random forest method. Validate the model with Iris_test.csv data?

	Iris-setosa	Iris-versicolor	Iris-virginica	class.error
Iris-setosa	50	0	0	0
Iris-versicolor	0	47	3	0.06
Iris-virginica	0	3	47	0.06

CLASSIFICATION AND REGRESSION TREE

Random Forest

Example

Develop a model to predict plant class using the data given in Iris.csv.\ using random forest method. Validate the model with Iris_test.csv data?

Model Validation

Read data to new data

```
> newdata <- read.csv("E:/Infosys/Part 2/Data/Iris_test.csv")  
> pred = predict(mymodel, newdata = newdata)  
> mytable = table(newdata$Class, pred)  
> mytable
```


CLASSIFICATION AND REGRESSION TREE

Random Forest

Example

Develop a model to predict plant class using the data given in Iris.csv.\ using random forest method. Validate the model with Iris_test.csv data?

	Predicted		
Actual	Iris-setosa	Iris-versicolor	Iris-virginica
Iris-setosa	49	0	0
Iris-versicolor	0	15	0
Iris-virginica	0	0	2

CLASSIFICATION AND REGRESSION TREE

Random Forest

Example

Develop a model to predict plant class using the data given in Iris.csv.\ using random forest method. Validate the model with Iris_test.csv data?

	Predicted		
Actual	Iris-setosa	Iris-versicolor	Iris-virginica
Iris-setosa	74.24	0.00	0.00
Iris-versicolor	0.00	22.73	0.00
Iris-virginica	0.00	0.00	3.03

**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.

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

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 $L(\text{value}) = 0$ when value is on target
2. Choose k such that loss $L(\text{value}) = 1$, when value is on specification limits

Taguchi's Loss Function Approach

Taguchi's Loss Function

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

1. Smaller the better type

$$\text{Target} = 0, k = 1 / \text{USL}^2$$

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

2. Larger the better type

$$\text{Target} = \infty, k = 1 / \text{LSL}^2$$

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

3. Nominal the best type

$$\text{Target} = t, k = 4 / (\text{USL} - \text{LSL})^2$$

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

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

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 $\text{Value}_{\text{minimum}}$

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

$d = 0$, if value $> \text{USL}$

$d = 1$. 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 $Value_{\text{maximum}}$

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

$d = 0$, if value $\leq LSL$

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

$Value_{\text{maximum}}$ 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, \dots, d_p , then

Overall Desirability

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

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

Exercise: Rate the clusters based on the following parameters using both the approaches

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 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

$\{\text{Milk, Bread}\} \rightarrow \{\text{Biscuits}\}$ 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

$\{\text{Milk, Bread, Biscuits}\} = 2$

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

$$\{\text{Milk, Bread, Biscuits}\} = 2 / 5$$

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

$$= \{\text{Milk, Biscuits, Toys}\} / \{\text{Milk, Biscuits}\} = 2 / 3 = 0.67$$

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 \geq minimum support

2. Rule Generation

Fix minimum confidence value

Generate high confidence rules from each frequent itemset

MARKET BASKET ANALYSIS

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

MARKET BASKET ANALYSIS

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

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, 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
$A \rightarrow C$	0.50	0.75
$A \rightarrow E$	0.50	0.75
$B \rightarrow E$	0.50	1.00
$C \rightarrow E$	0.50	0.75
$C \rightarrow A$	0.50	0.75
$E \rightarrow A$	0.50	0.60
$E \rightarrow B$	0.50	0.60
$E \rightarrow C$	0.50	0.60

MARKET BASKET ANALYSIS

Association Rule Mining: Other Measures

Lift

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

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 : $\text{Lift} \geq 1$

$\text{Lift}(A, C) = 1.12 > \text{Lift}(A, E)$ 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 file to my data and specify the variables

```
>target = mydata$items
```

```
>ident = mydata$id
```

Make transaction variable

```
>transactions = as(split(target, ident), "transactions")
```

Generate Rules

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

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

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

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 the 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

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		

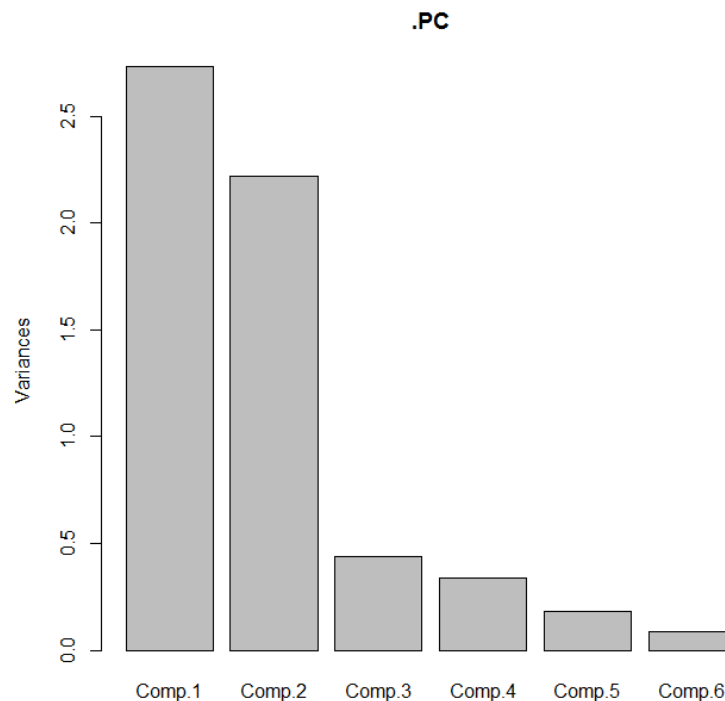
Number of factors selected : 2

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

FACTOR ANALYSIS

Step 5: Calculate Factor Scores– Eigen Vectors

>loadings(mymodel)

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

	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

FACTOR ANALYSIS**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

FACTOR ANALYSIS

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

FACTOR ANALYSIS

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

FACTOR ANALYSIS

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 of 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

CLUSTER ANALYSIS

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

R Code

Read data to mydata and compute distance

```
> distance = dist(mydata, method = "euclidean")
```

Generate Clusters

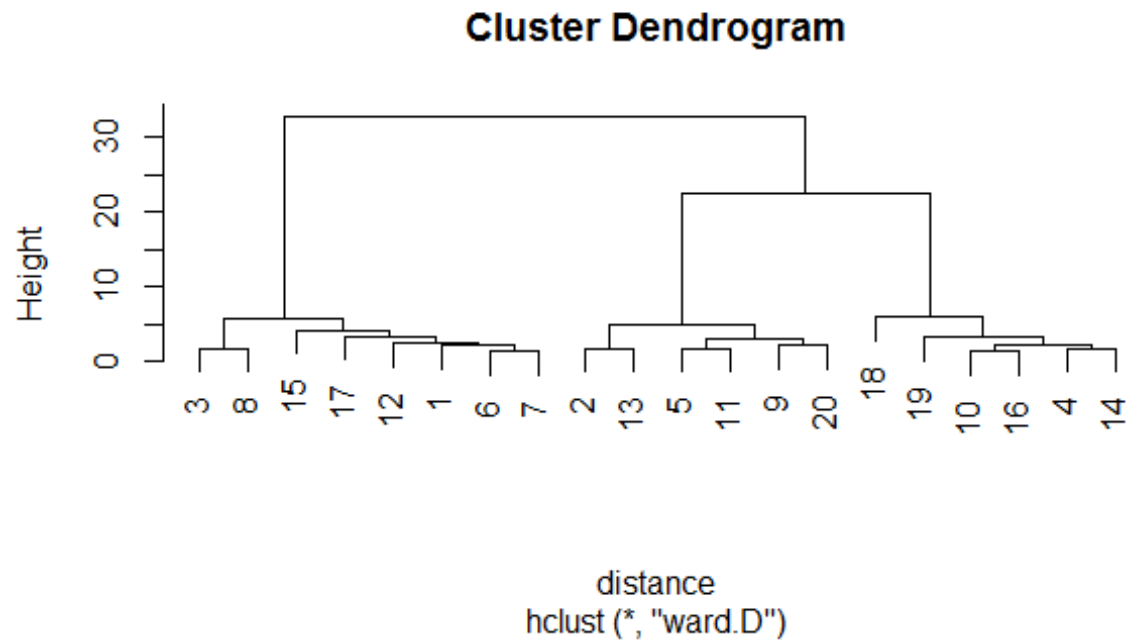
```
> mymodel = hclust(distance, method = "ward")
```

Plot Dendrogram

```
> plot(mymodel)
```

CLUSTER ANALYSIS

Decide on number of clusters: Dendrogram



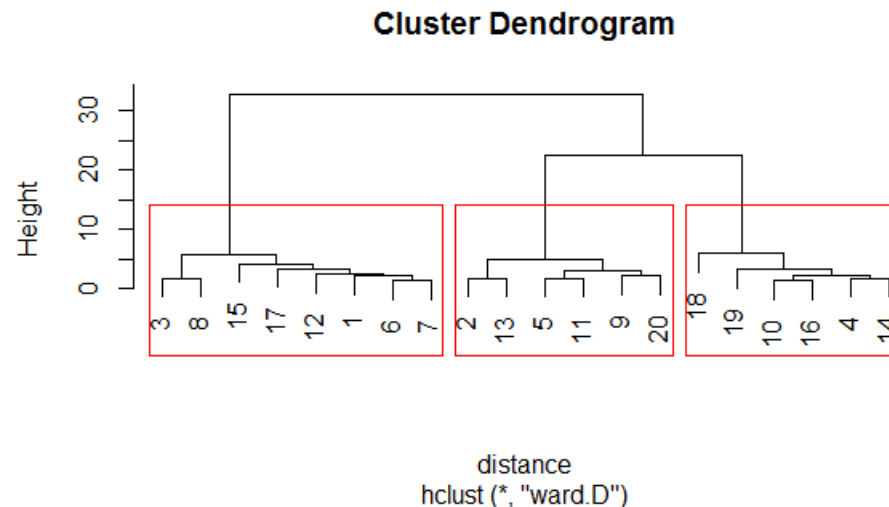
CLUSTER ANALYSIS

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")
```



CLUSTER ANALYSIS

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 ANALYSIS

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

- 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
- Assigns each observation to the most likely class, given its predictor values
- Uses the conditional probability of $P(y / x)$ for making prediction

Methodology

Assign a test observation with predictor vector x_0 to the class j for which

$$P(y = j / x = x_0)$$

is the largest

NAÏVE BAYES CLASSIFIER

Example : The data on code review duration and defect density obtained for 10 code reviews are given below. Predict the defect density for review duration = Low using Naïve Bayes classifier?

SL No	Review Duration	Defect Density
1	Low	High
2	Low	Medium
3	Low	Low
4	Low	Medium
5	Low	Medium
6	Low	High
7	High	Low
8	High	High
9	High	Low
10	High	High
11	High	Low
12	High	Low

Predict y given $x = \text{Low}$

NAÏVE BAYES CLASSIFIER

Example : The data on code review duration and defect density obtained for 10 code reviews are given below. Predict the defect density for review duration = Low using Naïve Bayes classifier?

SL No	Review Duration	Defect Density
1	Low	High
2	Low	Medium
3	Low	Low
4	Low	Medium
5	Low	Medium
6	Low	High
7	High	Low
8	High	High
9	High	Low
10	High	High
11	High	Low
12	High	Low

$$P(y = \text{Low} / x = \text{Low})$$

$$= \text{Number of cases when both } x \text{ \& } y \text{ are Low} / \text{Number of cases } x \text{ is Low} = 1/6 = 0.17$$

$$P(y = \text{Medium} / x = \text{Low})$$

$$= \text{Number of cases both } x \text{ is Low and } y \text{ is Medium} / \text{Number of cases } x \text{ is Low} = 3/6 = 0.50$$

$$P(y = \text{High} / x = \text{Low})$$

$$= \text{Number of cases both } x \text{ is Low and } y \text{ is High} / \text{Number of cases } x \text{ is Low} = 2/6 = 0.33$$

NAÏVE BAYES CLASSIFIER

Example : The data on code review duration and defect density obtained for 10 code reviews are given below. Predict the defect density for review duration = Low using Naïve Bayes classifier?

SL No	Review Duration	Defect Density
1	Low	High
2	Low	Medium
3	Low	Low
4	Low	Medium
5	Low	Medium
6	Low	High
7	High	Low
8	High	High
9	High	Low
10	High	High
11	High	Low
12	High	Low

Maximum of

$P(y = \text{Low} / x = \text{Low})$, $P(y = \text{Medium} / x = \text{Low})$ and $P(y = \text{High} / x = \text{Low})$

$\max(0.17, 0.50, 0.33) = 0.50$

$= P(y = \text{Medium} / x = \text{Low})$

Predicted value of y for $x = \text{Low}$ is $y = \text{Medium}$

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 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
 `> library(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

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

```
> pretest
```

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



Contact email : sqcbombay@gmail.com
sarkar.ashok@gmail.com