Lecture 1

Introduction to Statistics

-Dr. Umesh R A

Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

Agenda of Lecture 1

- Why Statistics?
- Introduction to statistics
- Branches of Statistics
- Terminology
- Data
- Road Map
- Statistical Survey
- Frequency Distribution
- Diagrammatic and graphical representation

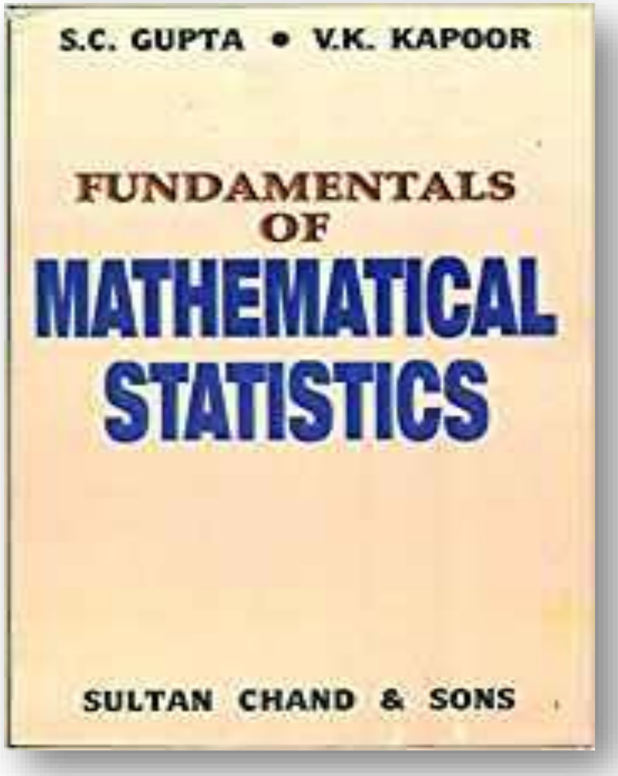
Learning Outcome

This session mainly designed to introduce you with fundamentals of statistics.

After this class, you can identify;

- Types of Data.
- You will know different terminology used during Stat/ML course.
- You can able to use and explain suitable Graphs/Diagrams.

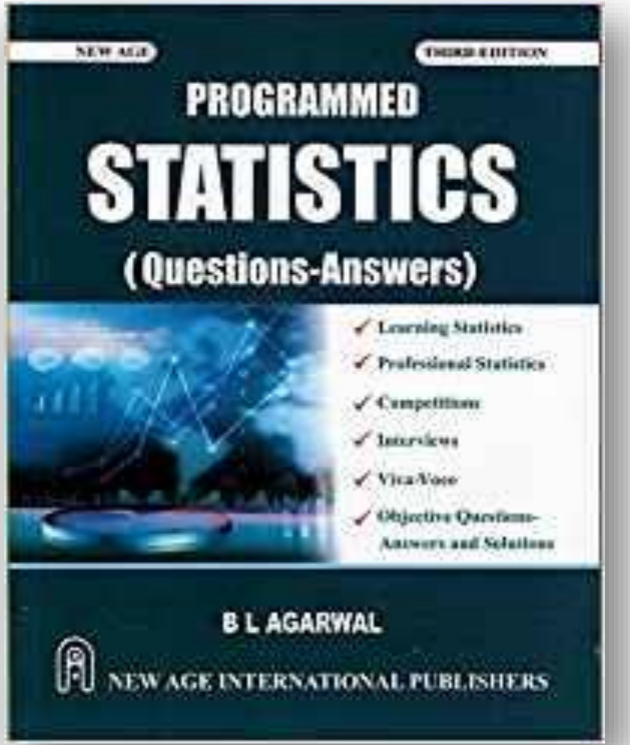
Reference Books



S.C. GUPTA • V.K. KAPOOR

FUNDAMENTALS OF MATHEMATICAL STATISTICS

SULTAN CHAND & SONS



NEW AGE

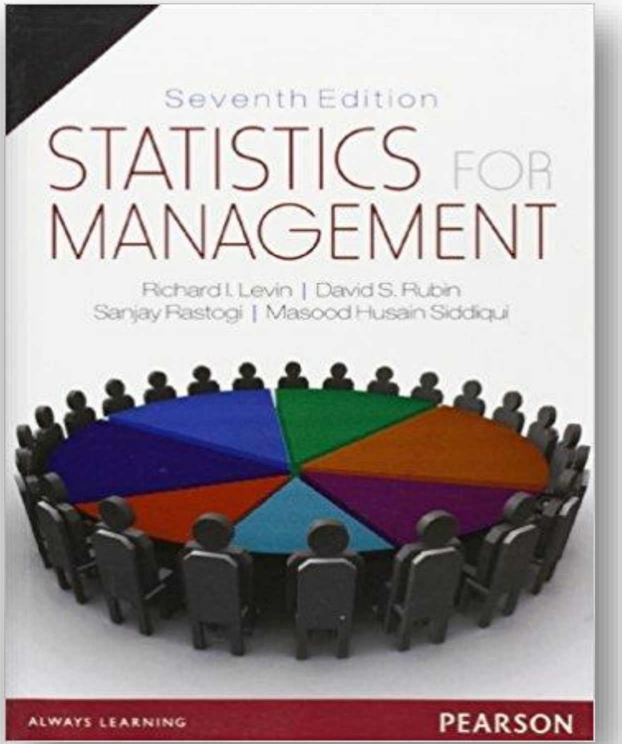
THIRD EDITION

PROGRAMMED STATISTICS (Questions-Answers)

✓ Learning Statistics
✓ Professional Statistics
✓ Competitions
✓ Interviews
✓ Viva-Voce
✓ Objective Questions-Answers and Solutions

B L AGARWAL

NEW AGE INTERNATIONAL PUBLISHERS



Seventh Edition

STATISTICS FOR MANAGEMENT

Richard I. Levin | David S. Rubin
Sanjay Rastogi | Masood Husain Siddiqui

ALWAYS LEARNING

PEARSON

Why Statistics?

- Data are everywhere!
- Statistical techniques are used to make many decisions that affect our lives.
- No matter what your career, you will make professional decisions based on data (information).
- An understanding of statistical methods will help you to make these decisions effectively.

Famous quote:

> *"In god we all trust, all others must bring data."*

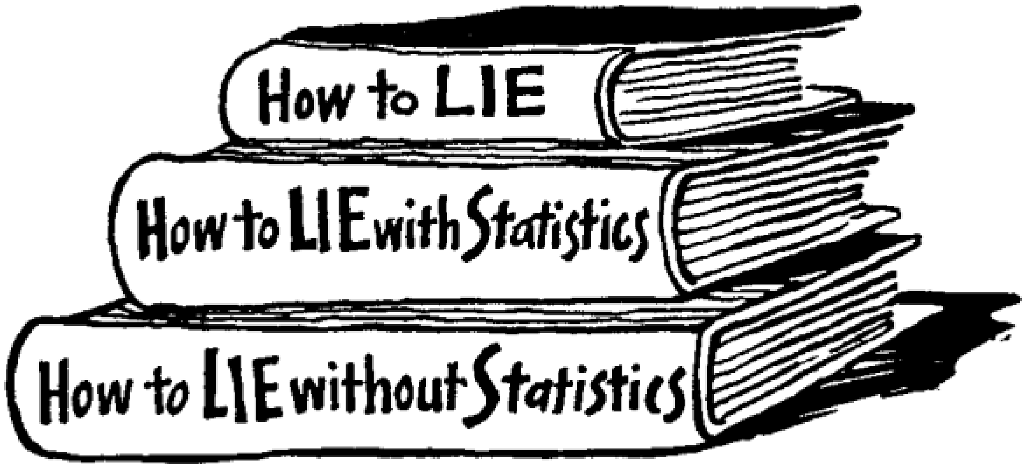
-W. Edward Deming

Distrust in Statistics

There are three kinds of lies:

“**lies**, **damned lies**, **and Statistics**.”

(-B.Disraeli)



Branches of Statistics

1. Theoretical Statistics (Statistical Methods)
2. Applied Statistics

Statistical Methods

- **Statistical methods** are those procedures used in the collection, presentation, analysis and interpretation of data.

Applied Statistics

- Biostatistics
- Time Series
- Reliability Theory
- Epidemiology
- Demography
- Design of Experiment
- Actuarial Statistics

- Statistical Quality Control
- Operation Research
  - LPP
  - Assignment Problem
  - Transportation Problem
  - Inventory Problem
  - Replacement Theory
  - Queuing Theory
- etc…

Meaning of word: “Statistics”

Meanings;

Plural sense

Collection of numerical fact.

Singular sense

Science of studying statistical methods.

- **Statistics** is the mathematical science that is involved with the collection, presentation, analysis and interpretation of data.
- **Statistician** is one who’s intention is to use sample information to make an inference about a population.

Statistical Investigation (Statistical Survey)

•Stages of Statistical survey

1. Planning and Preparation

- Purpose
- Scope of the survey
- Preparation of Frame
- Type of the survey
- Type of data collected
- Statistical units to be used
- Degree of accuracy
- Statistical methods

2. Execution of the survey

- Collection of Data
- Scrutiny, editing and presentation
- Analysis
- Interpretation

Important terms:

- Pilot Survey
- Census
- Sample Survey
- Sampling
- Sampling Frame
- Primary Data
- Secondary Data
- Sampling error
- Non-sampling Error

Population:

- A population is the totality of the observations.
- A population may be finite (small or large) or infinite.
- The characteristics of a population are called parameters

Sample:

- A sample is a subset of a population.
- A sample is usually of smaller size.
- The characteristics of a sample are called statistics.

Data

- Bases of Classification
  - –Qualitative,
  - –Quantitative,
  - –Spatial/Geographical,
  - –Temporal/Chronological/time

- Types of Classification
  - One-way Classification
  - Two-way Classification
  - Multi-way Classification

Flavours of Data

- Qualitative / Categorical / Attribute
- Quantitative / Variable-
  - Discrete
  - Continuous

Data (based on Scale)

- Nominal
- Ordinal
- Interval
- Ratio

Wake up call

Q.1. Following data belongs to which data scale?

Ticket No: 22A5, 38A7, 41A62, 45A8,…

Ans:

A. Nominal Scale

B. Ordinal Scale

C. Interval Scale

D. Ratio Scale

Wake up call

Q.1. Following data belongs to which data scale?

Ticket No: 22A5, 38A7, 41A62, 45A8,…

Ans:

A. Nominal Scale

B. Ordinal Scale

C. Interval Scale

D. Ratio Scale

Wake up call

Q.2. Following data belongs to which data scale?

Weight of fruit (in gms): 42, 38, 41, 45,…

Ans:

A. Nominal Scale

B. Ordinal Scale

C. Interval Scale

D. Ratio Scale

Wake up call

Q.2. Following data belongs to which data scale?

Weight of fruit (in gms): 42, 38, 41, 45,…

Ans:

A. Nominal Scale

B. Ordinal Scale

C. Interval Scale

D. Ratio Scale

Different Terminologies

- **Independent variable** is called a predictor variable, controlled variable, explanatory variable, etc.

- **Dependent variable** is sometimes called a response variable, predicted variable, explained variable, output variable, etc.

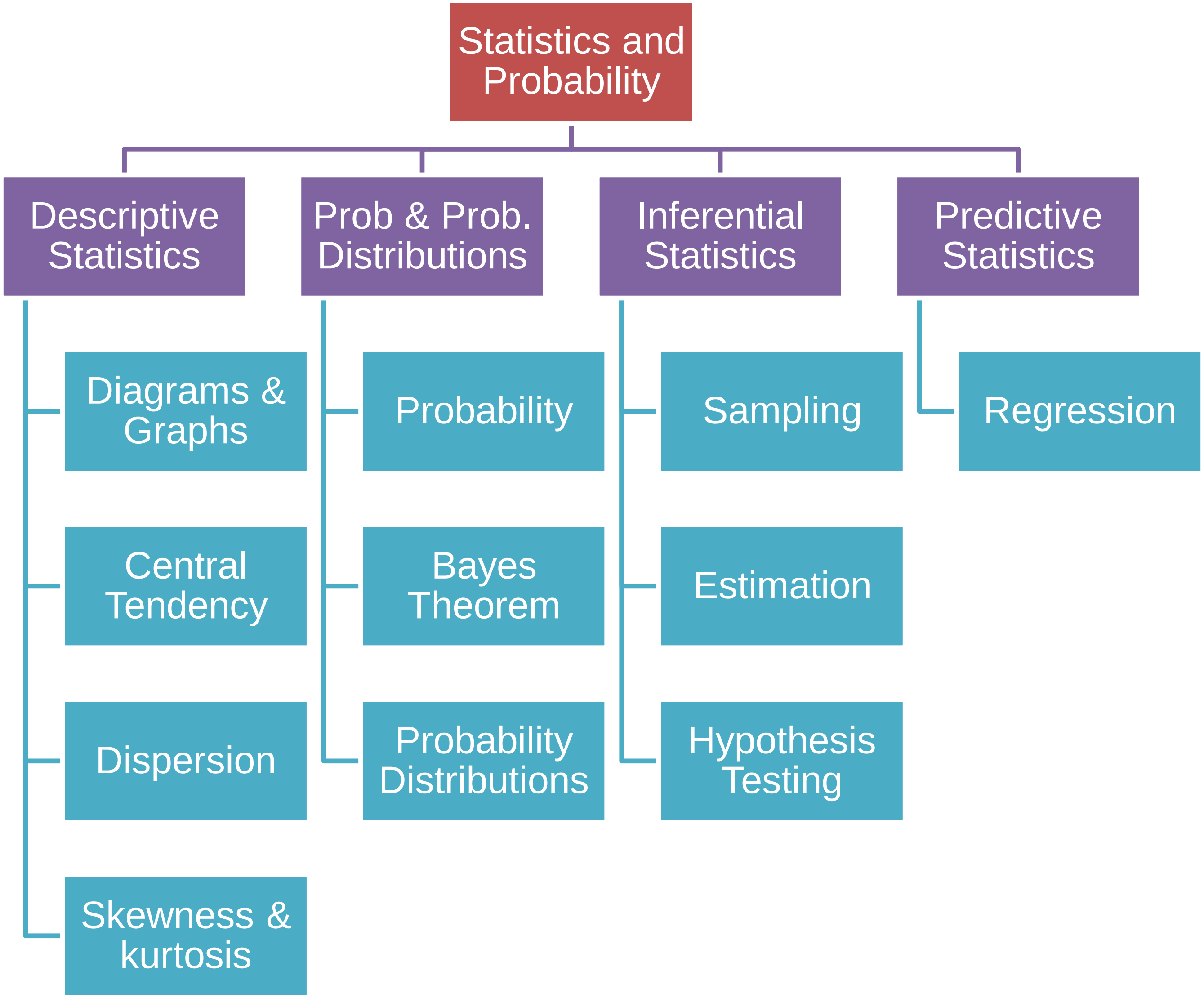
Remember !

All you need to start your Journey is;

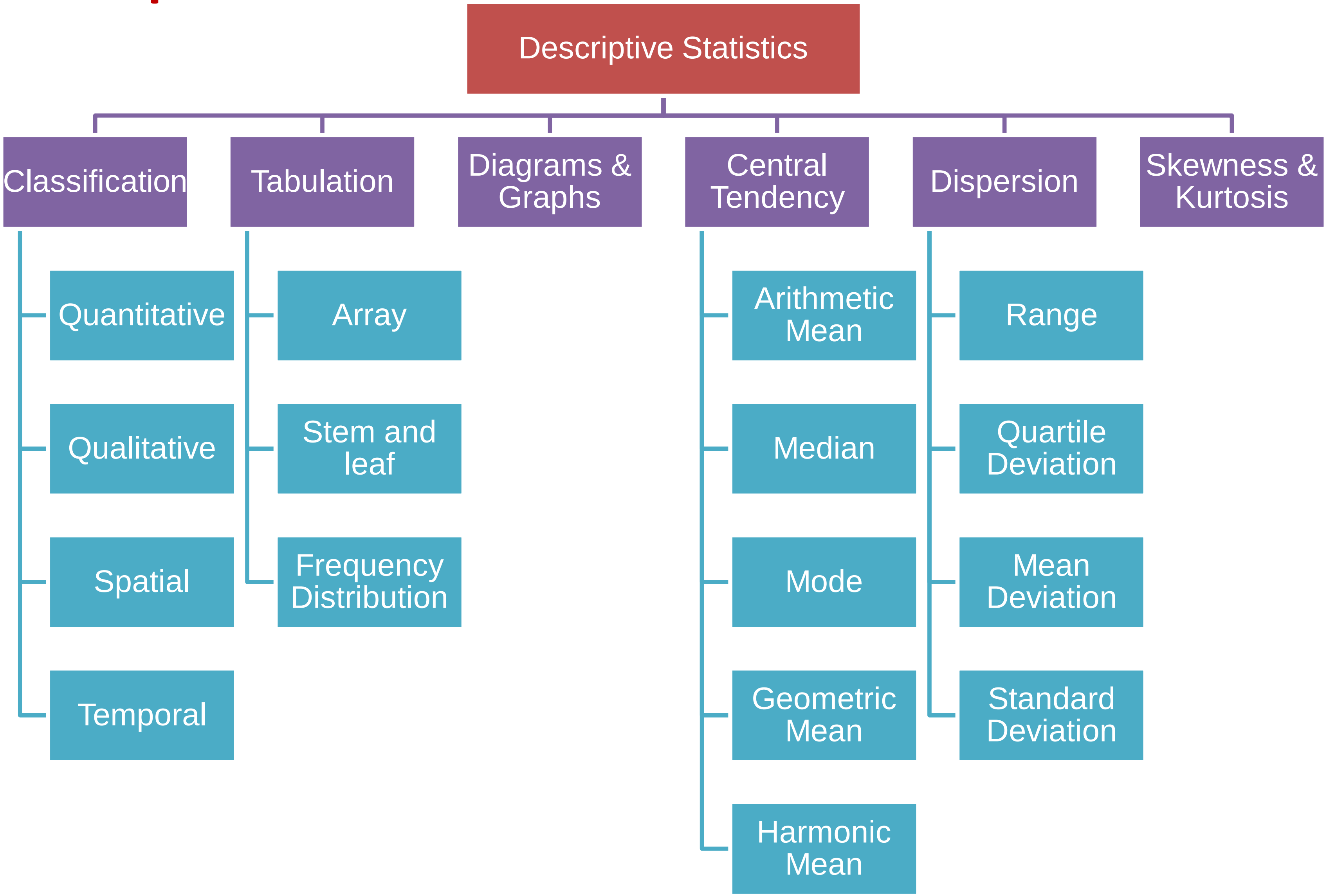
1. **Plan**
2. **Road Map**
3. **Determination**

Road Map

Road Map

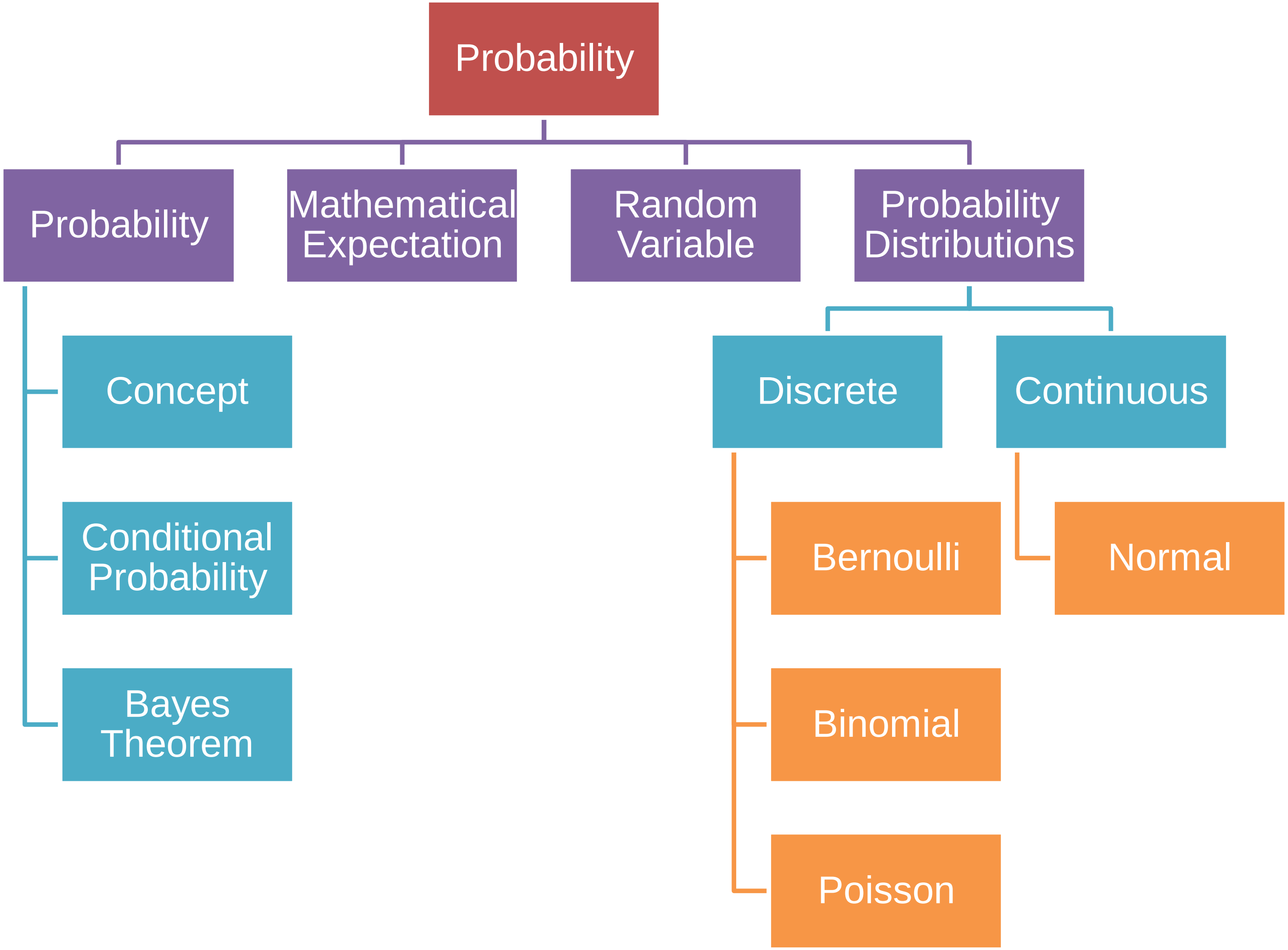

Statistics and Probability
Descriptive Statistics
Prob & Prob. Distributions
Inferential Statistics
Predictive Statistics
Diagrams & Graphs
Central Tendency
Dispersion
Skewness & kurtosis
Probability
Bayes Theorem
Probability Distributions
Sampling
Estimation
Hypothesis Testing
Regression

Road Map



- Descriptive Statistics
  - Classification
    - Quantitative
    - Qualitative
    - Spatial
    - Temporal
  - Tabulation
    - Array
    - Stem and leaf
    - Frequency Distribution
  - Diagrams & Graphs
  - Central Tendency
    - Arithmetic Mean
    - Median
    - Mode
    - Geometric Mean
    - Harmonic Mean
  - Dispersion
    - Range
    - Quartile Deviation
    - Mean Deviation
    - Standard Deviation
  - Skewness & Kurtosis

Road Map



Probability

Probability

Mathematical Expectation

Random Variable

Probability Distributions

Concept

Conditional Probability

Bayes Theorem

Discrete

Continuous

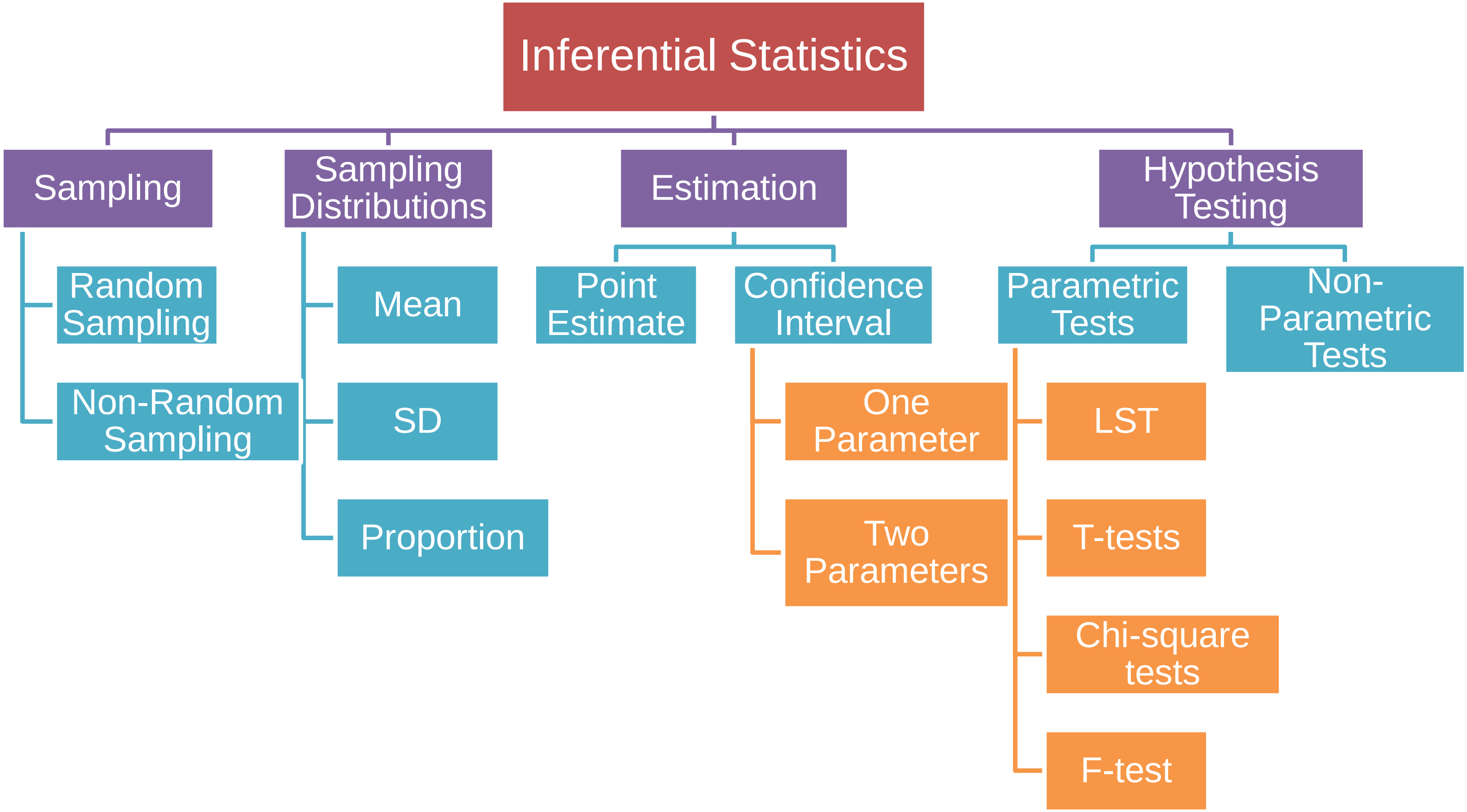
Bernoulli

Binomial

Poisson

Normal

Road Map


Inferential Statistics

Sampling

Random Sampling

Non-Random Sampling

Sampling Distributions

Mean

SD

Proportion

Estimation

Point Estimate

Confidence Interval

One Parameter

Two Parameters

Hypothesis Testing

Parametric Tests

LST

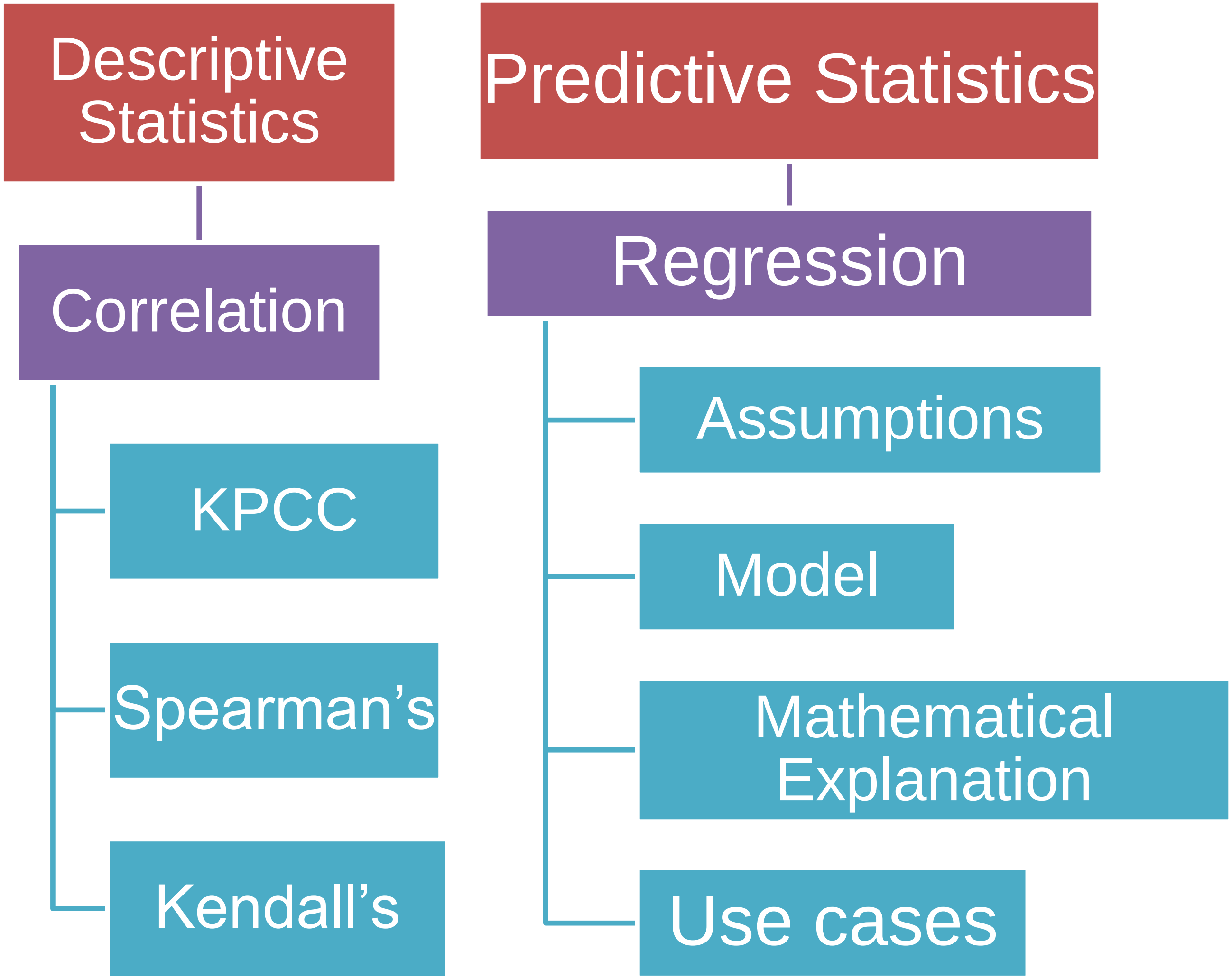
T-tests

Chi-square tests

F-test

Non-Parametric Tests

Road Map



Descriptive Statistics

Correlation

KPCC

Spearman's

Kendall's

Predictive Statistics

Regression

Assumptions

Model

Mathematical Explanation

Use cases

Descriptive Statistics

Statistical Representation of Data

- Massive volume of statistical data (raw *or* unorganized data). **(Situation)**
- Difficult to examining and interpret the unorganized data. **(Problem)**
- Therefore, it should be organized. **(Solution)**

- Tools: **Classification**, **Tabulation** and **Graphic representation**

Classification and Tabulation

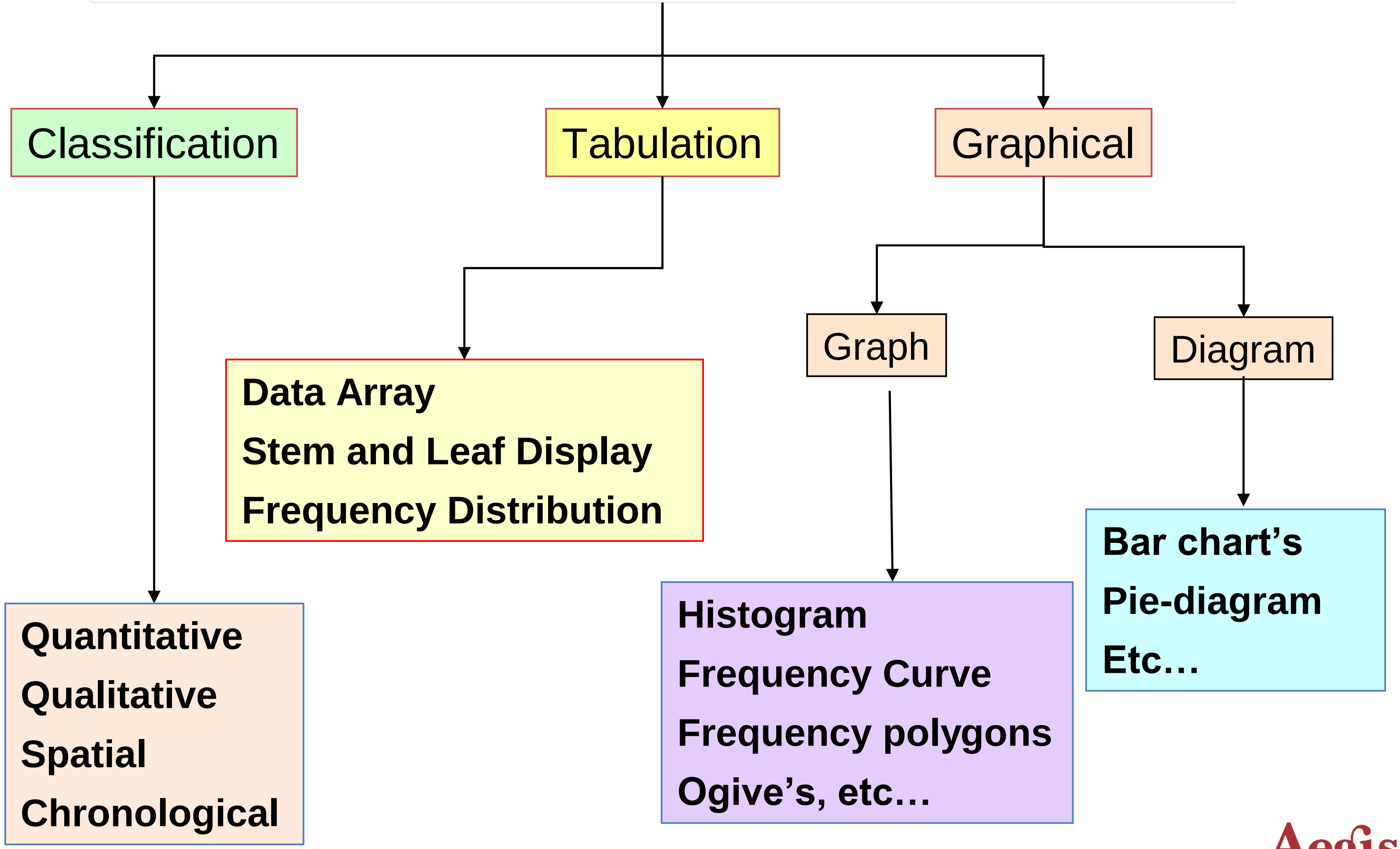
Classification

- Grouping of data according to their common characteristics.

Tabulation

- It is a **systematic** presentation of classified data in columns and rows.
- This sort of logical arrangement makes the data **easy to understand**, **facilitates comparisons** and provides an **effective way** of convey information to a reader.

Techniques for Data Representation



Classification

- Quantitative
- Qualitative
- Spatial
- Chronological

Tabulation

- Data Array
- Stem and Leaf Display
- Frequency Distribution

Graphical

Graph

- Histogram
- Frequency Curve
- Frequency polygons
- Ogive’s, etc…

Diagram

- Bar chart’s
- Pie-diagram
- Etc…

Data Array:

- simplest ways to present data.
- It arranges values in ascending or descending order.
- A minimum, maximum and repeated values are easily determined.

Data Array

67	72	75	85	88	89	89	90	99	100

Frequency Distribution:

A frequency distribution is a tabular arrangement of data whereby the data is grouped into different intervals, and then the number of observations that belong to each interval is determined.

C.I.	Frequency
0 – 2	20
3 – 5	14
6 – 8	15
9 – 11	2
12 – 14	1
Total	52

Stem-and-Leaf Display:

- A clear disadvantage of frequency table is that the identity of individual observations is lost in grouping process.
- To overcome this drawback, John Tukey (1977) introduced this technique.

Stem	Leaves
6	7
7	2 5
8	5 8 9 9
9	0 4 9

Guidelines for Frequency Tables

1. Be sure that the classes are not **overlapping**.
2. Include all classes, even if the **frequency is 0**.
3. Try to use the same width for all classes.
4. Select **convenient numbers** for class limits.
5. Use between **5** and **20** classes.
6. The sum of the class frequencies must **equal** the number of original data values.

Constructing a Frequency Table

1. Decide on the **number** of classes .

2. Determine the class width by dividing the range by the number of classes (range = highest score - lowest score) and round up

$$\text{class width} \quad \approx \quad \text{round up of} \; \frac{\text{range}}{\text{number of classes}}$$

3. Select for the **first lower limit** either the lowest score or a convenient value slightly less than the lowest score.

[cont…]

4. Add the **class width** to the starting point to get the second lower class limit, add the width to the second lower limit to get the third, and so on.
5. List the lower class limits in a vertical column and enter the upper class limits.
6. Represent each score in the appropriate class and count them to find the total frequency for each class.

Sturge's Rule

Used to determining the desirable number of classes/groups.

$$\boldsymbol{K = 1 + [3.2 \times log(N)]}$$

K- No. of classes

N- No. of observations

Example:

Make a frequency distribution from the following set of measurements for a particular sample:

2.5	5.9	3.2	1.4	7.0	4.3	8.9	0.7	4.2	9.9
3.4	4.6	5.0	6.4	1.1	9.2	7.7	0.9	4.0	2.3
5.6	2.2	3.1	4.7	5.5	6.6	1.9	3.9	6.1	5.2
8.2	3.3	2.2	5.8	4.1	3.8	1.2	6.8	9.5	0.8

Solution:

- By scanning the data, we find that the
  - Minimum value = 0.7 and Maximum value = 9.9
  - **Range** = 9.9 - 0.7 = **9.2**
- Suppose we decide to take 10 classes, then
  - Size or width of equal class interval = 9.2/10 = 0.92
  - So use **h = 1**
- Therefore, we can create the following **10** classes.

Raw Discrete Data

2	2	5	1	2	6	3	3	4	2
4	0	5	7	7	5	6	6	8	10
7	2	2	10	5	8	2	5	4	2
6	2	6	1	7	2	7	2	3	8
1	5	2	5	2	14	2	2	6	3
1	7								

C.I.	Frequency
0 – 2	20
3 – 5	14
6 – 8	15
9 – 11	2
12 – 14	1
Total	**52**

Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

Lower Class Limits

are the smallest numbers that can actually belong to different classes

Lower Class Limits

C.I.	Frequency
0 - 2	20
3 - 5	14
6 - 8	15
9 - 11	2
12 - 14	1

Upper Class Limits

are the smallest numbers that can actually belong to different classes

Upper Class Limits

C.I.	Frequency
0 - 2	20
3 - 5	14
6 - 8	15
9 - 11	2
12 - 14	1

Class Midpoints / Class Mark

are midpoints of the classes.

Class Mid-points

C.I.	Frequency
0 - 1 2	20
3 - 4 5	14
6 - 7 8	15
9 - 10 11	2
12 - 13 14	1

Class Width

It is the difference between two consecutive lower class limits or two consecutive class boundaries

Class Width

	C.I.	Frequency
3	0 - 2	20
3	3 - 5	14
3	6 - 8	15
3	9 - 11	2
3	12 - 14	1

Relative Frequency Table

C.I.	Frequency
0 - 2	20
3 - 5	14
6 - 8	15
9 - 11	2
12 - 14	1
Total	52

Total frequency = 52

C.I.	Relative Frequency
0 - 2	38.5%
3 - 5	26.9%
6 - 8	28.8%
9 - 11	3.8%
12 - 14	1.9%
	100

$$\text{Relative Frequency} = \frac{\text{Class Frequency}}{\text{Sum of all Frequencies}}$$

Frequency Density

- Note: Frequency density is useful when classes are of unequal widths.

$$\textbf{Frequency Density} = \frac{\textbf{Class Frequency}}{\textbf{Class Width}}$$

Cumulative Frequency Table

C.I.	Frequency
0 - 2	20
3 - 5	14
6 - 8	15
9 - 11	2
12 - 14	1

C.I.	Cumulative Frequency
Less than 3	20
Less than 6	34
Less than 9	49
Less than 12	51
Less than 15	52

Cumulative Frequencies

Frequency Tables

C.I.	Frequency
0 - 2	20
3 - 5	14
6 - 8	15
9 - 11	2
12 - 14	1

C.I.	Relative Frequency
0 - 2	38.5%
3 - 5	26.9%
6 - 8	28.8%
9 - 11	3.8%
12 - 14	1.9%

C.I.	Cumulative Frequency
Less than 3	20
Less than 6	34
Less than 9	49
Less than 12	51
Less than 15	52

Few more concepts

- Marginal Frequency Distribution
- Conditional Frequency Distribution

Wake up call

- Q.1 calculate frequency densities for the following data.

CI	frequency
0-30	60
30-40	60
40-60	20

Wake up call

- Q.1 calculate frequency densities for the following data.

CI	frequency	Frequency Density
0-30	60	2
30-40	60	6
40-60	20	1

Diagrammatic and Graphical representation

- Visual representations to be useful in **highlighting** information.

- **Graphs: histogram**, **frequency curve**, **frequency polygons, Ogive, etc**.

- **Diagram: Dot Plot, bar chart, Pie-diagrams, etc.**

Why Visual Representation?

- They are attractive
- They gives Birds eye view of the data
- They can be easily understood by common man
- They provides way of comparison of various characteristics
- Impression is long lasting

Diagrams

- One Dimensional:
  - Simple Bar Diagram
  - Multiple Bar Diagram
  - Subdivided Bar Diagram
  - Percentage Bar Diagram
- Two Dimensional
  - Pie diagram
- Three Dimensional

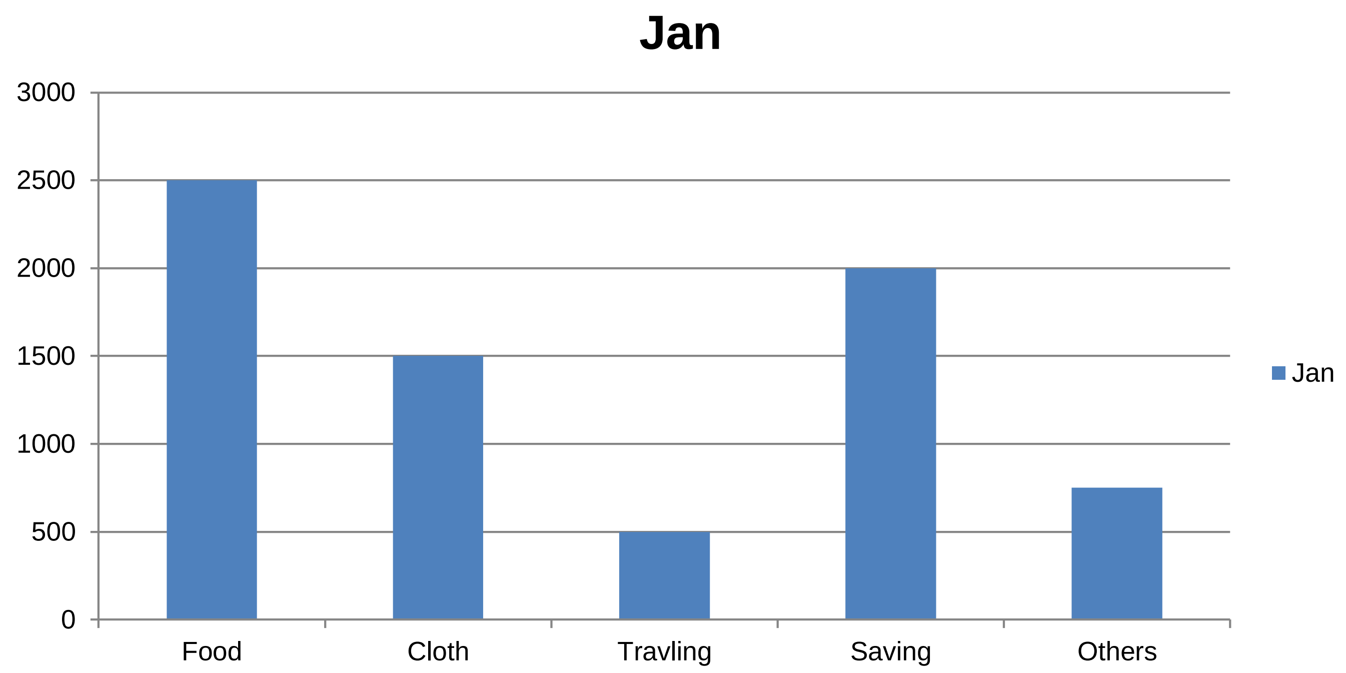
Graphs

- Histogram
- Frequency Curve
- Frequency Polygon
- Ogive
- Time series Graph (Historigram)
- Box plot
- Scatter Plot

Diagrams

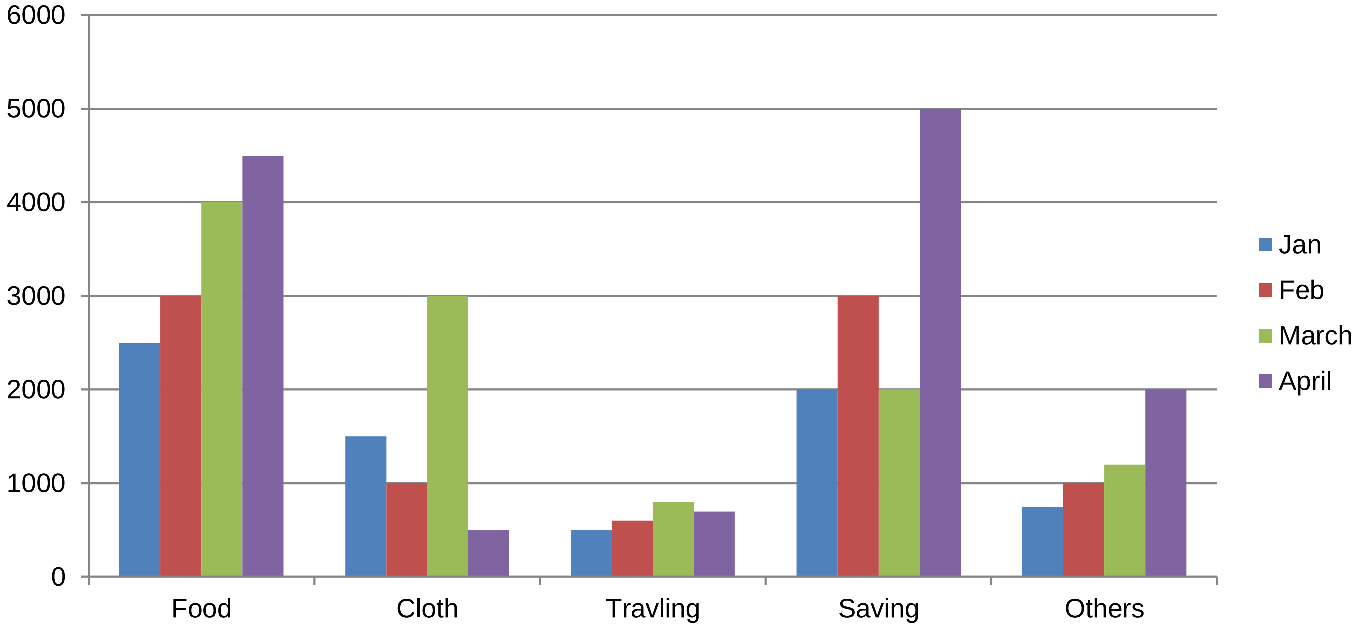
Bar Diagram

Items	Food	Cloth	Traveling	Savings	Others
Expenditure	2500	1500	500	2000	750



Multiple bar Diagram

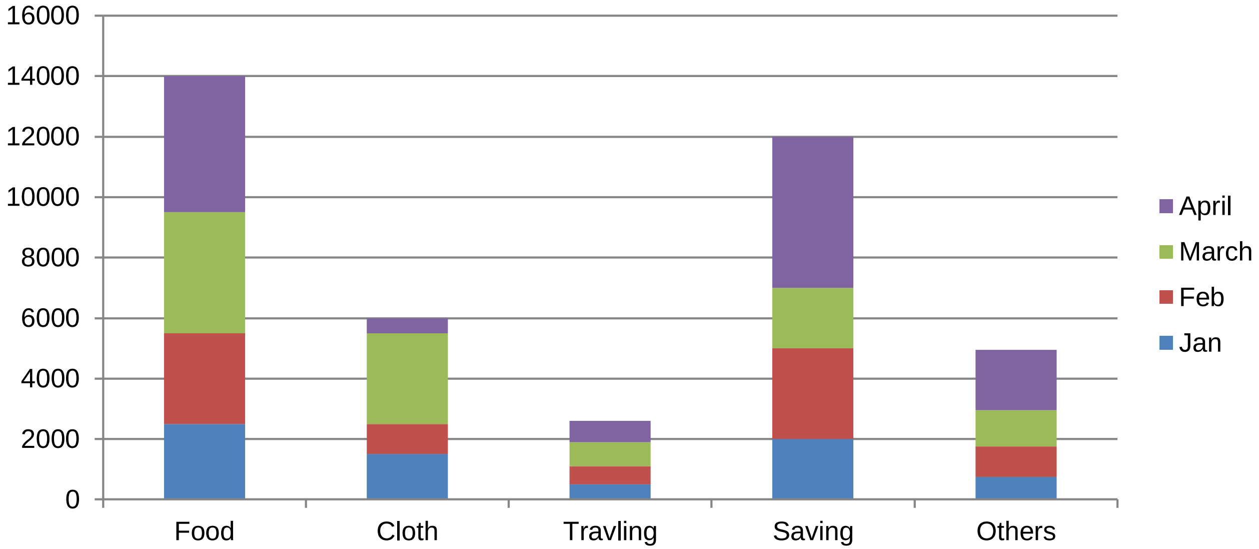
	Expenditure			
Items	Jan	Feb	March	April
Food	2500	3000	4000	4500
Cloth	1500	1000	3000	500
Travling	500	600	800	700
Savings	2000	3000	2000	5000
Others	750	1000	1200	2000



Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

Sub-divided Bar Diagram

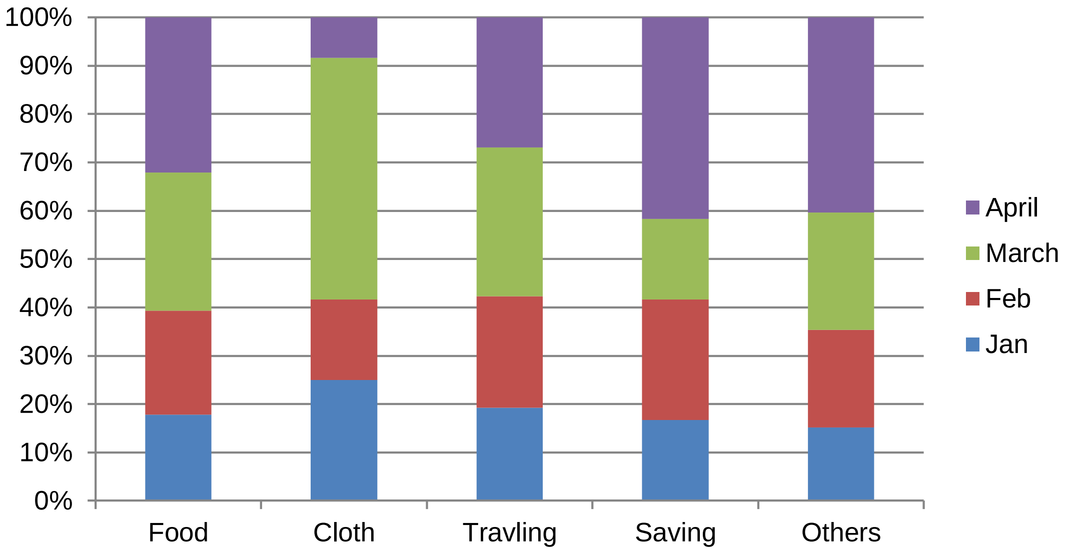
	Expenditure			
Items	Jan	Feb	March	April
Food	2500	3000	4000	4500
Cloth	1500	1000	3000	500
Travling	500	600	800	700
Savings	2000	3000	2000	5000
Others	750	1000	1200	2000



Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

Percentage Bar Diagram

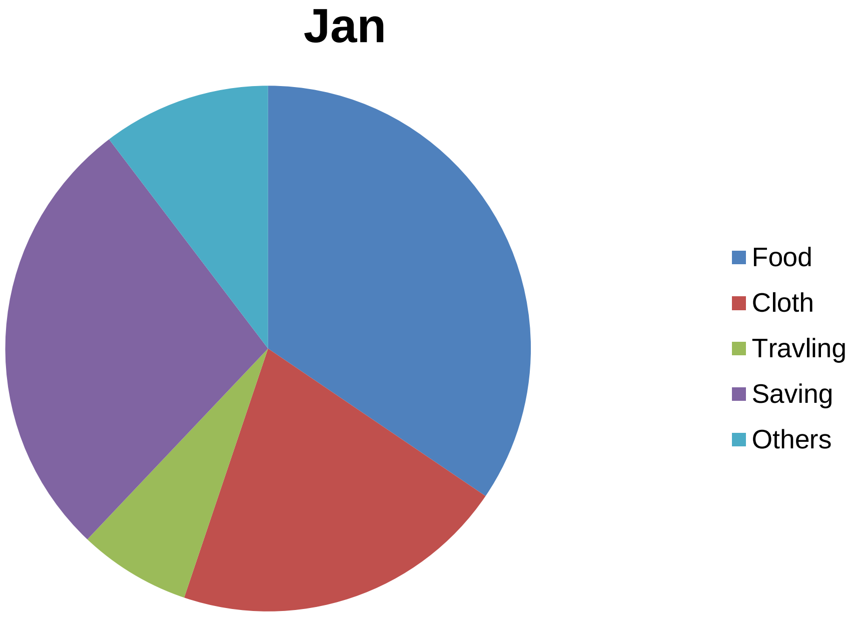
	Expenditure			
Items	Jan	Feb	March	April
Food	2500	3000	4000	4500
Cloth	1500	1000	3000	500
Travling	500	600	800	700
Savings	2000	3000	2000	5000
Others	750	1000	1200	2000



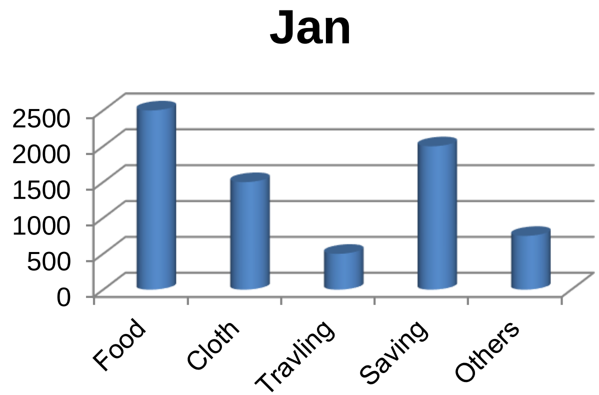
Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

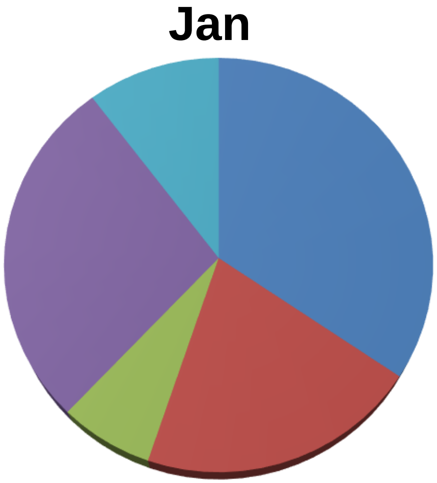
Pie Diagram

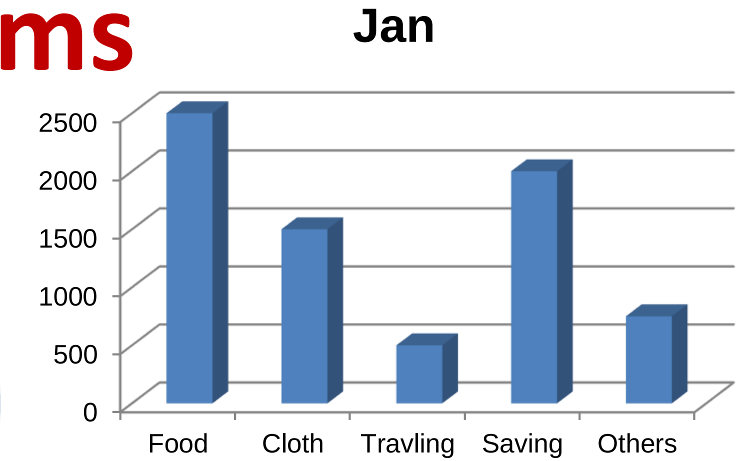
Items	Food	Cloth	Traveling	Savings	Others
Expenditure	2500	1500	500	2000	750

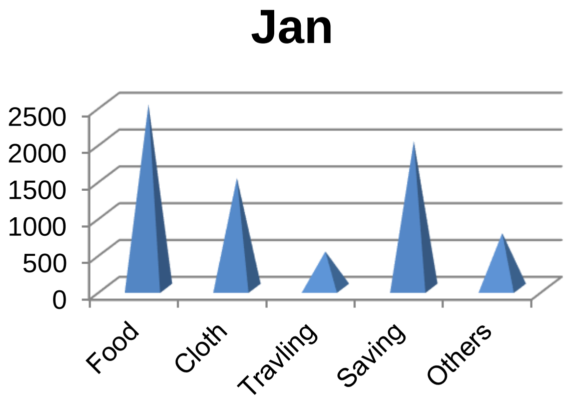


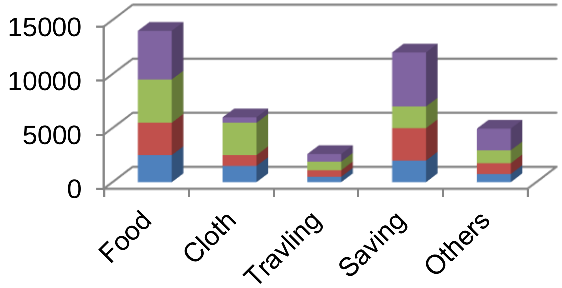
3D Diagrams

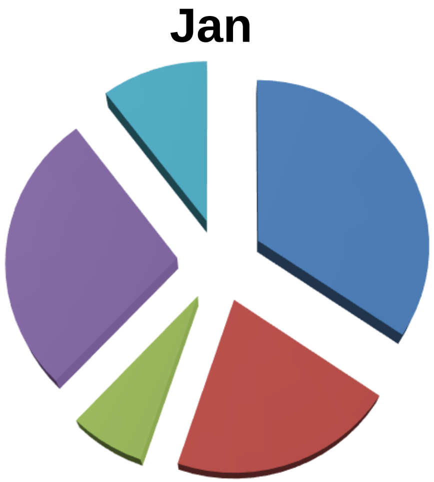


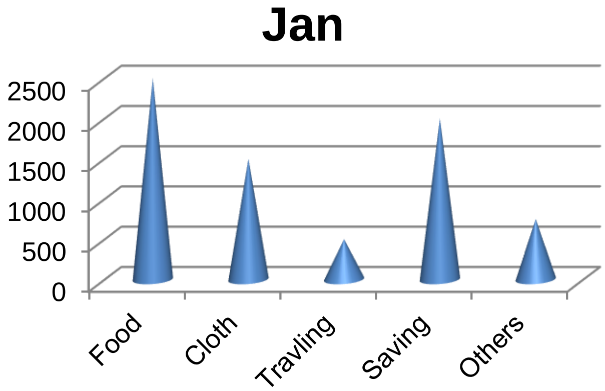


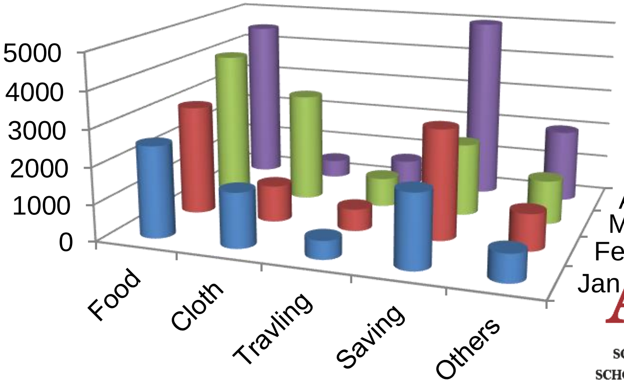










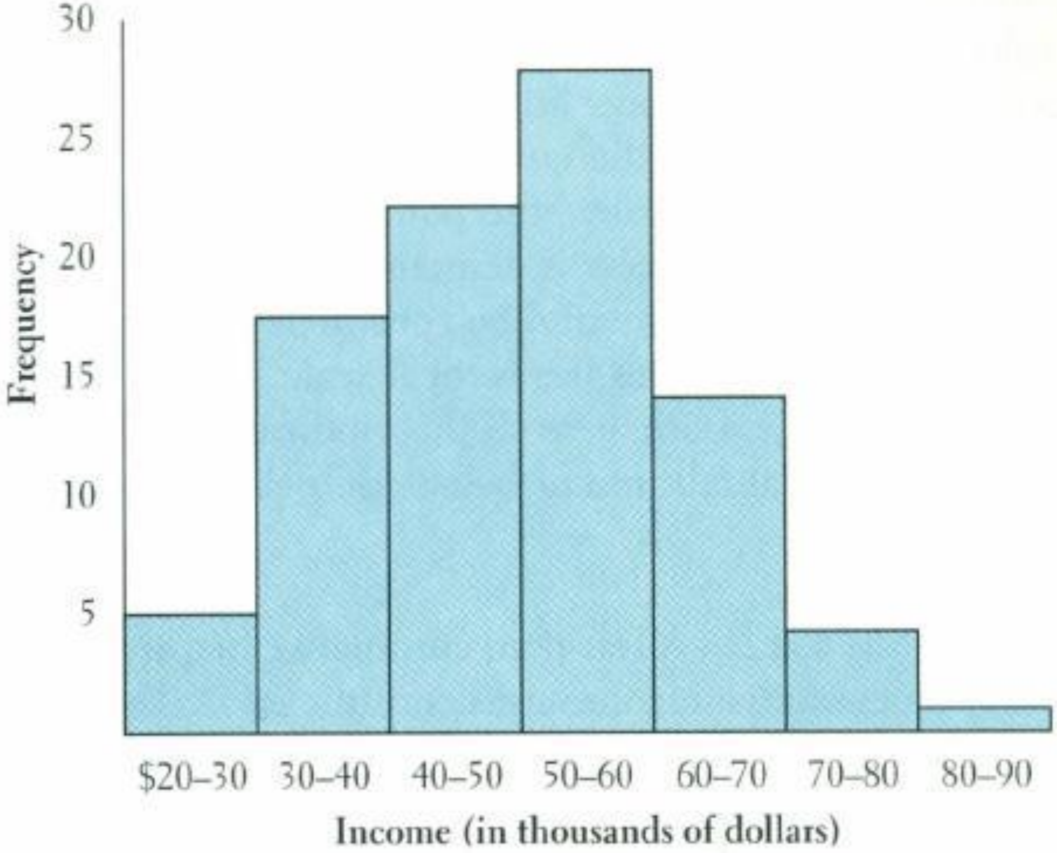


Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

Graphs

Histogram

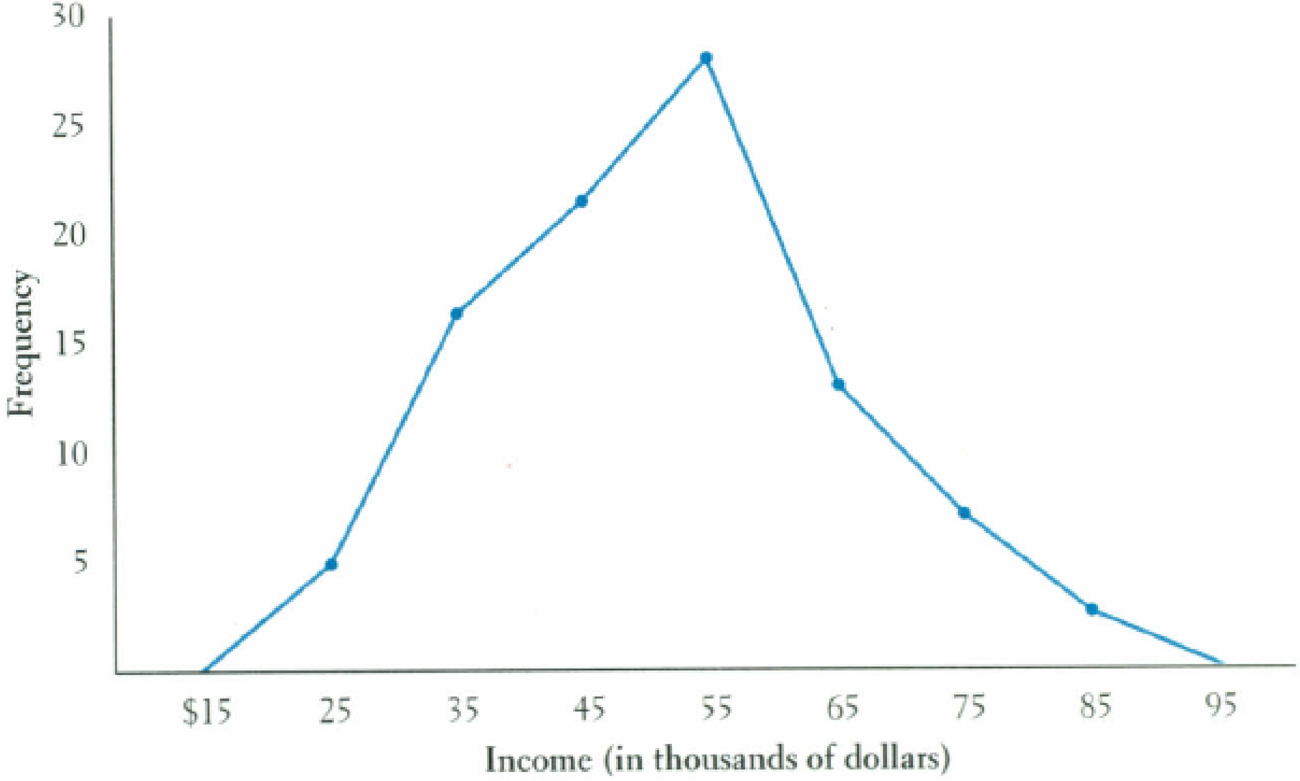
FIGURE 3.7 Histogram—Executive Incomes for the Sunrunner Corporation



Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

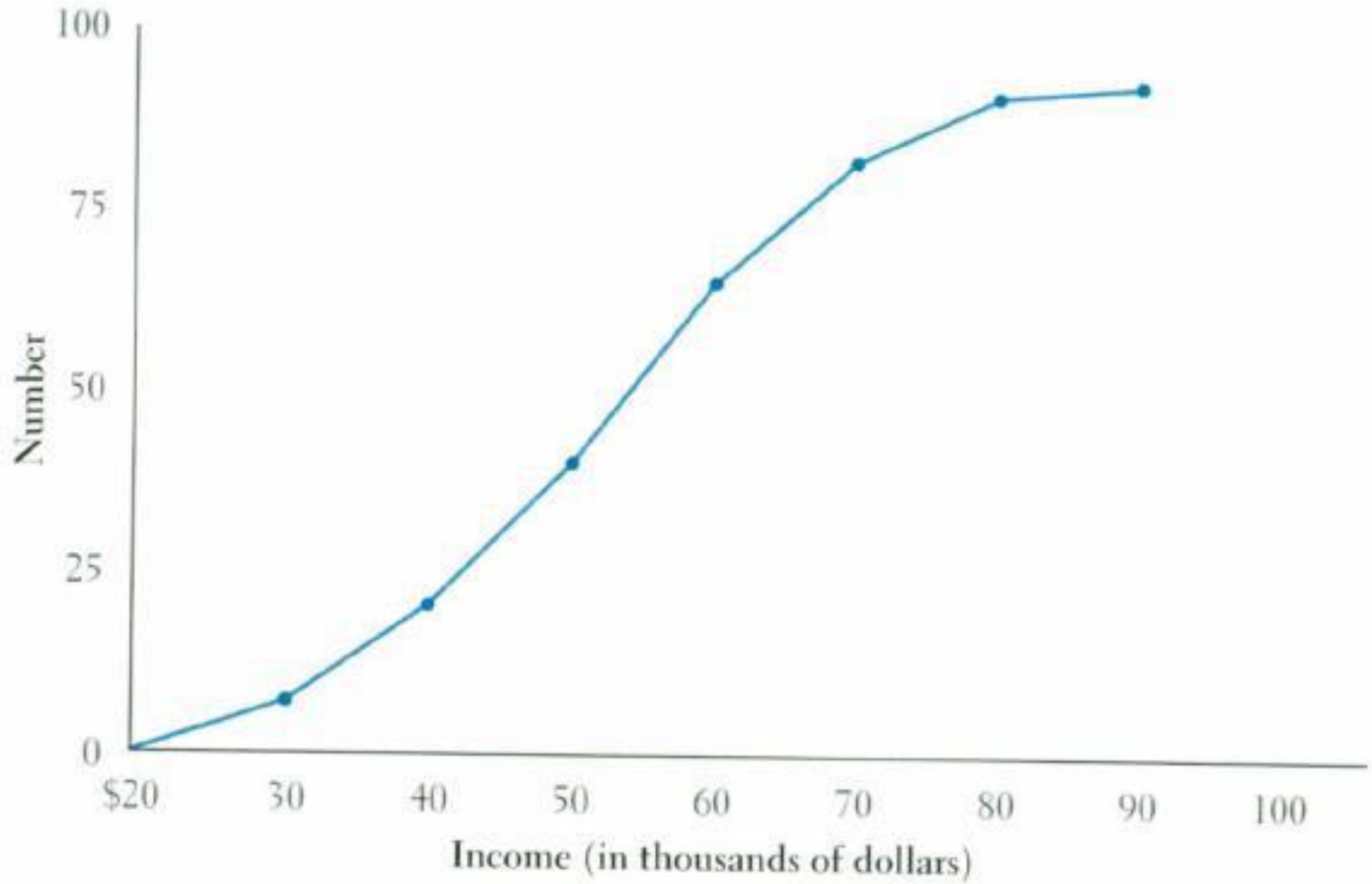
Frequency Polygon

FIGURE 3.8 Frequency Polygon—Executive Incomes



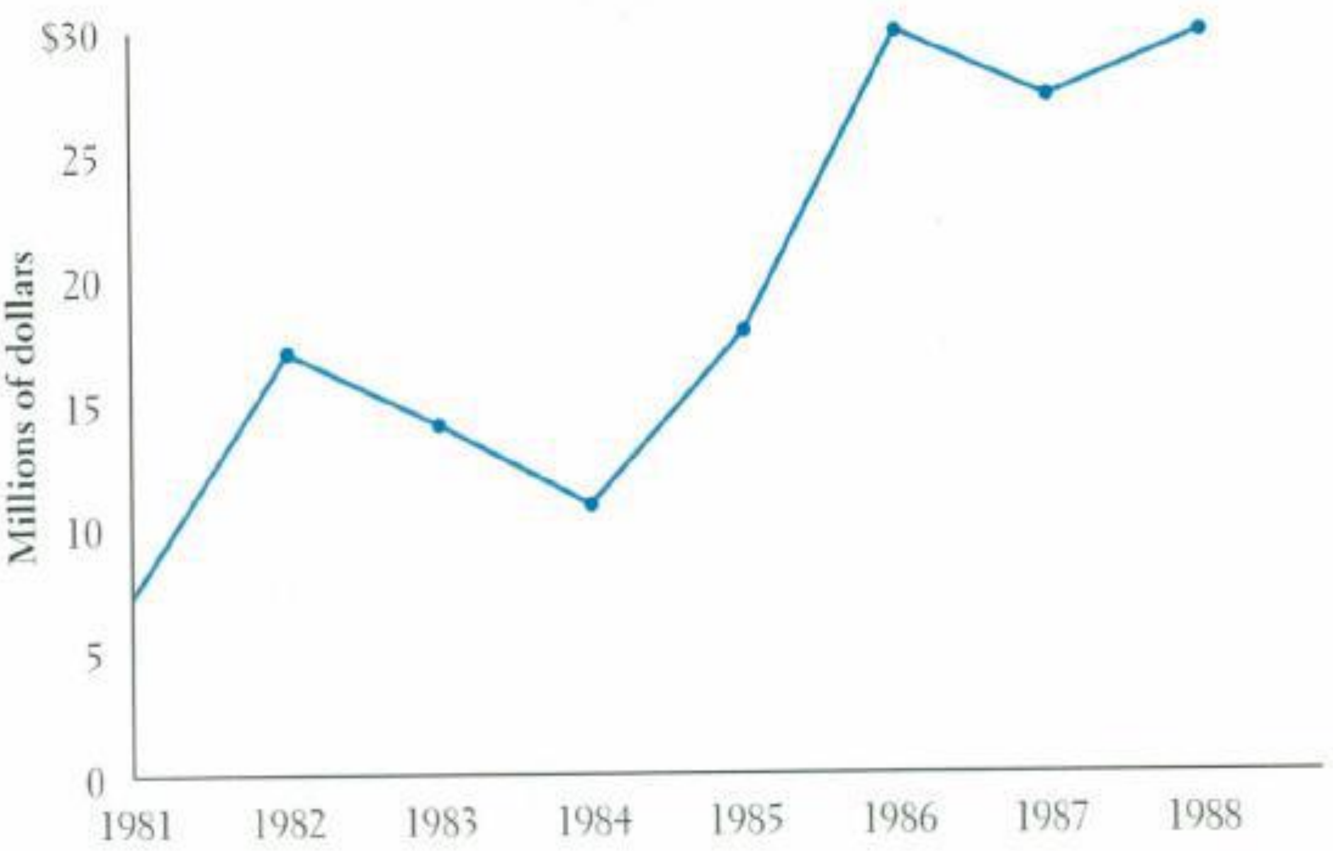
Less than Ogive

FIGURE 3.9 Ogive—Executive Incomes (frequencies)

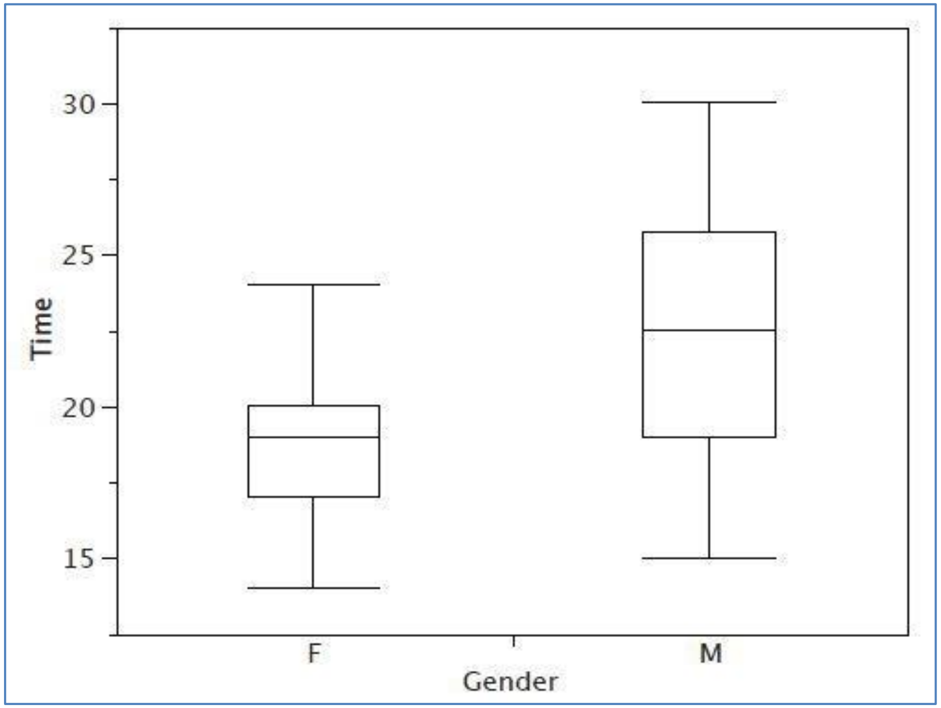


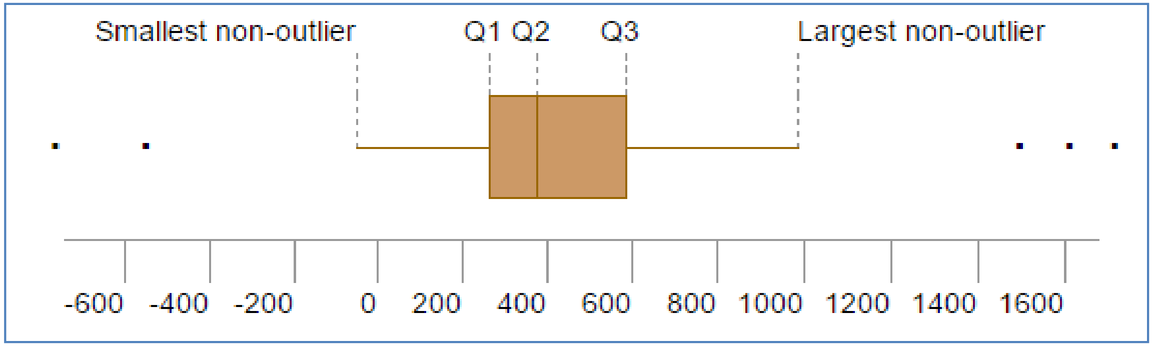
Historigram

FIGURE 3.13 Time Series Graph—Corporate Revenue, Flightcraft Corp.



Box Plot



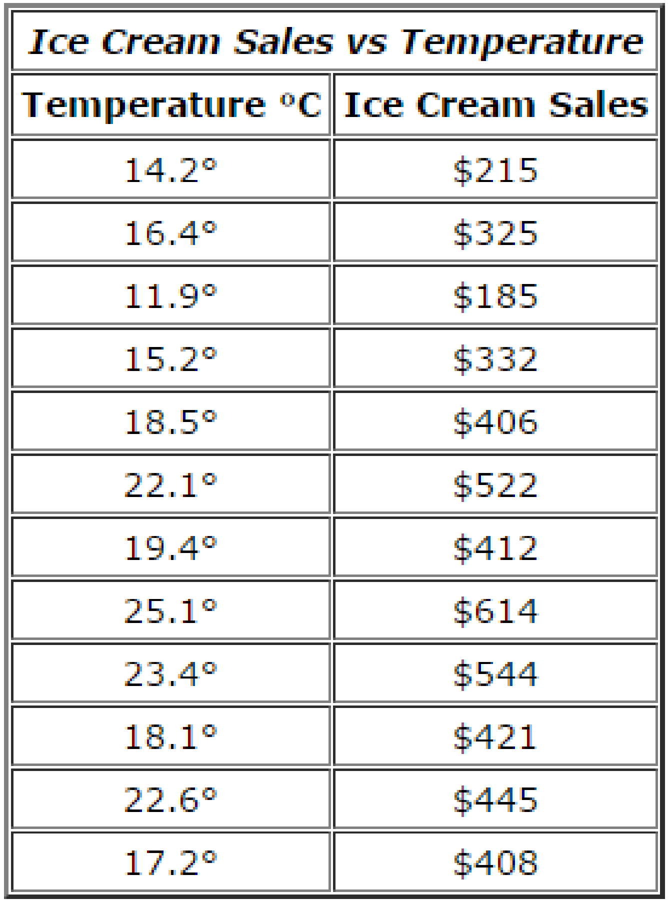


Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

Scatter Plot

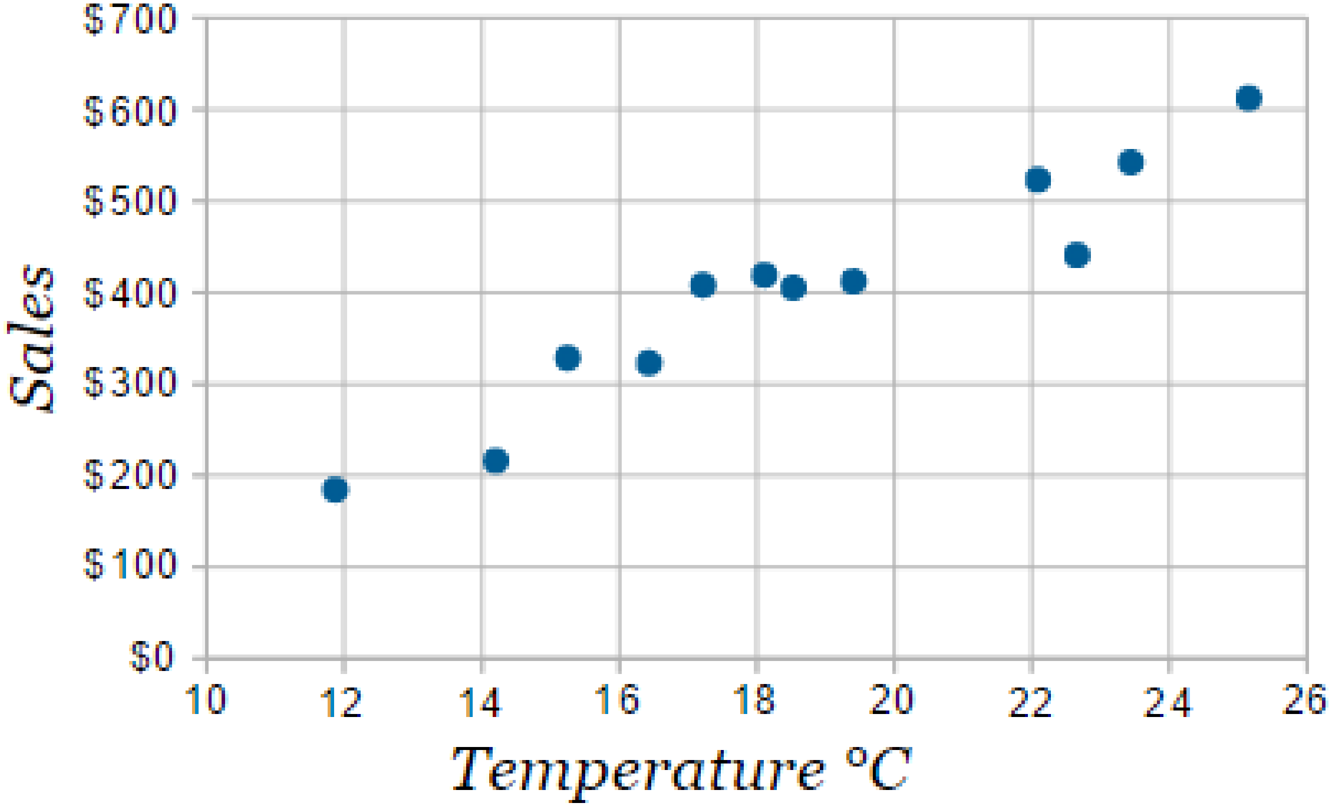
Example:

The local ice cream shop keeps track of how much ice cream they sell versus the noon temperature on that day. Here are their figures for the last 12 days:



Ice Cream Sales vs Temperature

Temperature °C	Ice Cream Sales
14.2°	$215
16.4°	$325
11.9°	$185
15.2°	$332
18.5°	$406
22.1°	$522
19.4°	$412
25.1°	$614
23.4°	$544
18.1°	$421
22.6°	$445
17.2°	$408


$700
$600
$500
$400
$300
$200
$100
$0
Sales
10
12
14
16
18
20
22
24
26
Temperature °C

Wake up call

Q.1. A Histogram contains a set of

A. Adjacent rectangles

B. Non Adjacent Rectangles

C. Adjacent squares

D. Adjacent triangles

Wake up call

Q.1. A Histogram contains a set of

A. Adjacent rectangles

B. Non Adjacent Rectangles

C. Adjacent squares

D. Adjacent triangles

Wake up call

Q.2. A circle in which sectors represents various quantities is called

A. Histogram

B. Frequency Polygon

C. Pie Chart

D. Component Bar chart

Wake up call

Q.2. A circle in which sectors represents various quantities is called

A. Histogram

B. Frequency Polygon

C. Pie Chart

D. Component Bar chart

Wake up call

Q.3. When data are arranged at regular interval of time, the classification is called:

A. Qualitative

B. Quantitative

C. Chronological

D. Geographical

Wake up call

Q.3. When data are arranged at regular interval of time, the classification is called:

A. Qualitative

B. Quantitative

C. Chronological

D. Geographical

Next Lecture

- Topic: **Descriptive Statistics I**
  - **Measures of Central Tendency (AM, Weighted Mean, GM, HM, Median, Mode)**
  - **Measures of Partition (Quartiles, Deciles, Percentiles)**
- Where you will find reference to study
  - Book: Statistics Class-11, Chapter1, pp.81-165
- Background material to study
  - Book: Business Statistics, Chapter1, pp.10-41
- We will have MCQ test on this lecture.
- And will have recap of the lecture 1.
- Discussion on Assignment 1 (If needed).