APPLIED STATISTICS

Section I

- Overview of statistics, Data description, frequency distributions, measures of central tendency and variability
- Probability and probability distributions: discrete and continuous random variables, moment generating function, expectation and standard deviation, joint distribution function, coefficient of correlation, and regression analysis.
- Standard probability distributions; Binomial, Poisson, Geometric, exponential and Gamma distribution, Normal, Weibull distributions.
- use of statistical package.

Section II

- Differentiation between a population and a sample, Sample with repetition, without repetition, sample mean, sample variance
- Student’s t and F- distributions, Chi square distributions
- Parameter estimation, confidence interval, level of significance, p-value
- Hypothesis testing, Type I and Type II errors, test for the population mean and variance
- Contingency table and the Chi-square test of independence
- F-test to compare the variances of two populations
- Use of statistical package

Books

- Probability and Statistics in Engineering

  William W. Hines, Douglas C. Montgomery, David M. Goldsman, Connie M. Borror

- Probability and Statistics

  SCHAUM ’ S Outline of Probability and Statistics

MURRAY R. SPIEGEL, JOHN SCHILLER, AND R. ALU SRINIVASAN

Statistics

Science of collection, presentation, analysis, and reasonable interpretation of collected information

presents a rigorous scientific method for gaining insight into data.

To make inference and predict relations of variables

Population

A set of similar items or events which of interest for some experiment.

Sample

A small sub-collection of population

Statistical model

A mathematical model that symbolizes a set of statistical assumptions about the generation of sample data

Statistical Description of Data

- Statistics describes a numeric set of data by its
  - Center
  - Variability
  - Shape

- Statistics describes a categorical set of data by
  - Frequency
  - percentage or proportion of each category

Variable - any characteristic of an individual or entity.

A variable can take different values for different individuals.

Variables can be *categorical* or *quantitative*.

Types of measurement

Discrete:

Quantitative data are called discrete if the sample space contains a finite or countably infinite number of values.

How many days did you watch a movie during the last 7 days?

Types of measurement

Continuous:

Quantitative data are called continuous if the sample space contains an interval or continuous span of real numbers.

e. g. Weight, height, temperature

– Height: 1.72 meters, 1.7233330 meters

- **Nominal**
  - ➢ Categorical variables with
  - ➢ No inherent order or ranking sequence

Value may be a numerical, but without numerical value

e. g. measurement such as female vs. male.

The only operation that can be applied to Nominal variables is enumeration.

- **Ordinal**

  Variables with an inherent rank or order e.g. mild, moderate, severe.

  Can be compared for equality, or greater than or less than, but not *how much* greater or less.

- **Interval**

  Values of the variable are ordered as in Ordinal, and additionally, differences between values are meaningful, however, the scale is not absolutely fixed.

  Calendar dates and temperatures on the Fahrenheit scale are examples.

  Addition and subtraction can be performed

  Multiplication and division are not meaningful operations.

- **Ratio** –

  Variables with all properties of Interval plus an absolute, non-arbitrary zero point,
  e.g. age, weight, temperature (Kelvin).

  Addition, subtraction, multiplication, and division are all meaningful operations.

Qualitative vs. Quantitative variables –

Qualitative variables:

values are texts

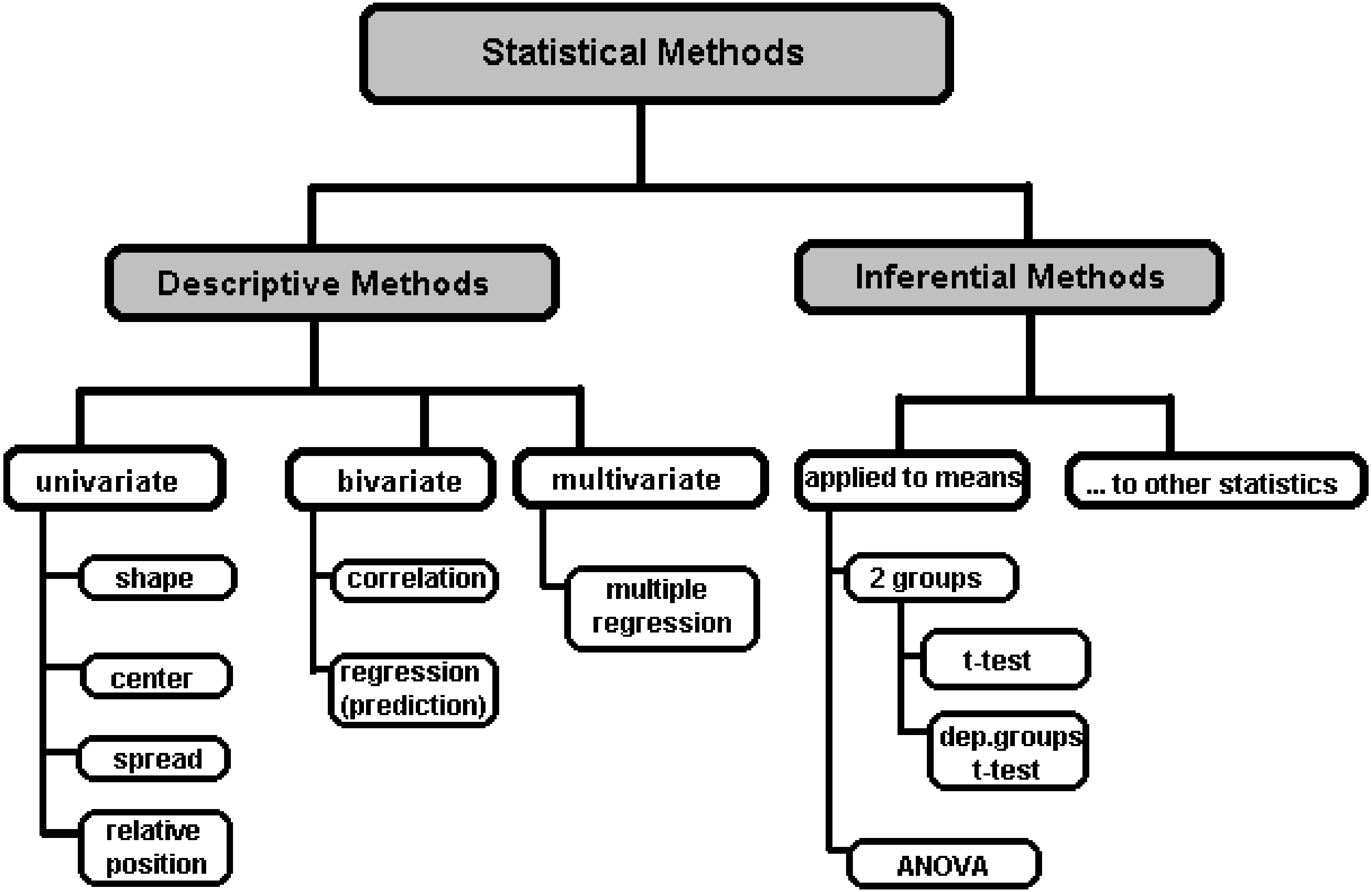
e.g. Female, male, boy, girl

Also called as string variables.

Quantitative variables:

numeric variables.

A Taxonomy of Statistics



Statistical Methods

Descriptive Methods

univariate

shape

center

spread

relative position

bivariate

correlation

regression (prediction)

multivariate

multiple regression

Inferential Methods

applied to means

2 groups

t-test

dep.groups t-test

ANOVA

... to other statistics

- ❖ Frequency Distribution

Consider a data set of 26 children of ages 1-6 years

Ungrouped Data

Age	**1**	**2**	**3**	**4**	**5**	**6**
Frequency	**5**	**3**	**7**	**5**	**4**	**2**

Ungrouped Data

Age Group	**1-2**	**3-4**	**5-6**
Frequency	**8**	**12**	**6**

- ❖ Cumulative Frequency

Data Presentation

Graphical Presentation:

We look for the overall pattern and for striking deviations from that pattern.

Over all pattern usually described by shape, center, and spread of the data.

An individual value that falls outside the overall pattern is called an *outlier*.

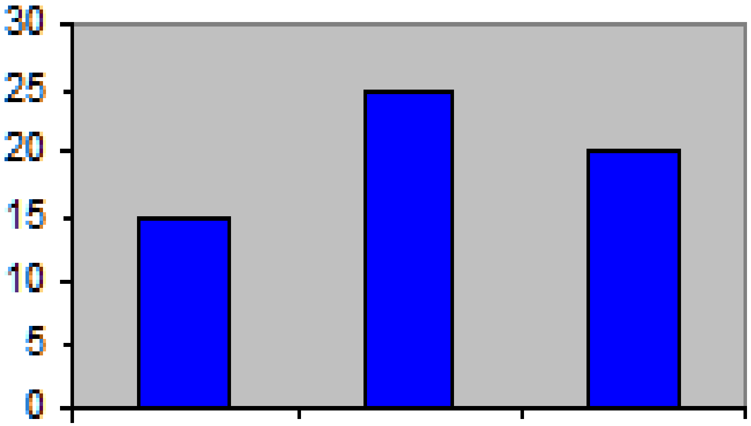
Bar diagram and Pie charts are used for categorical variables.

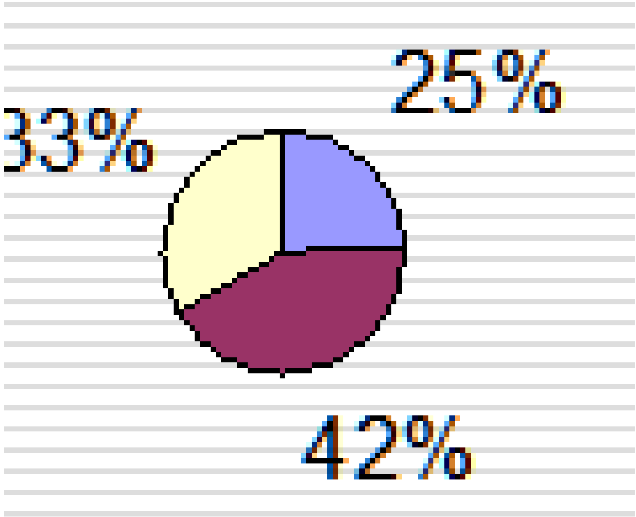
Histogram, stem and leaf and Box-plot are used for numerical variable.

Data Presentation –Categorical Variable

Bar Diagram: Lists the categories and presents the percent or count of individuals who fall in each category.

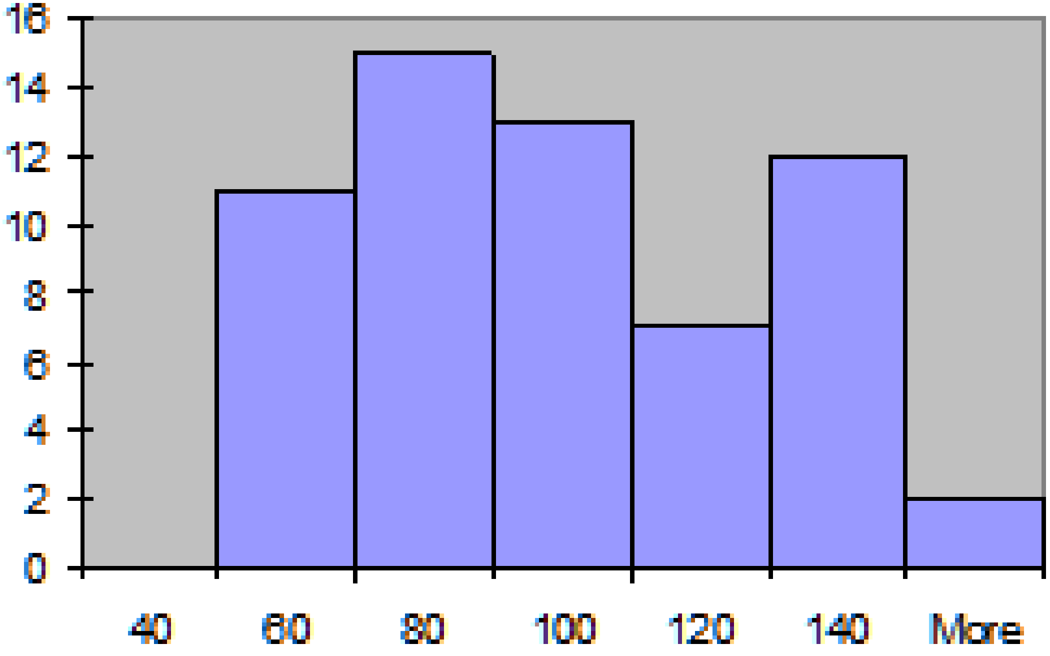
Pie Chart: Lists the categories and presents the percent or count of individuals who fall in each category.





Graphical Presentation –Numerical Variable

Histogram: Overall pattern can be described by its shape, center, and spread.



Box-Plot: Describes the five-number summary

Five Number Summary

The five number summary of a distribution consists of

- the smallest (Minimum) observation,
- the first quartile (Q1)
- the median(Q2)
- the third quartile
- the largest (Maximum) observation written in order from smallest to largest

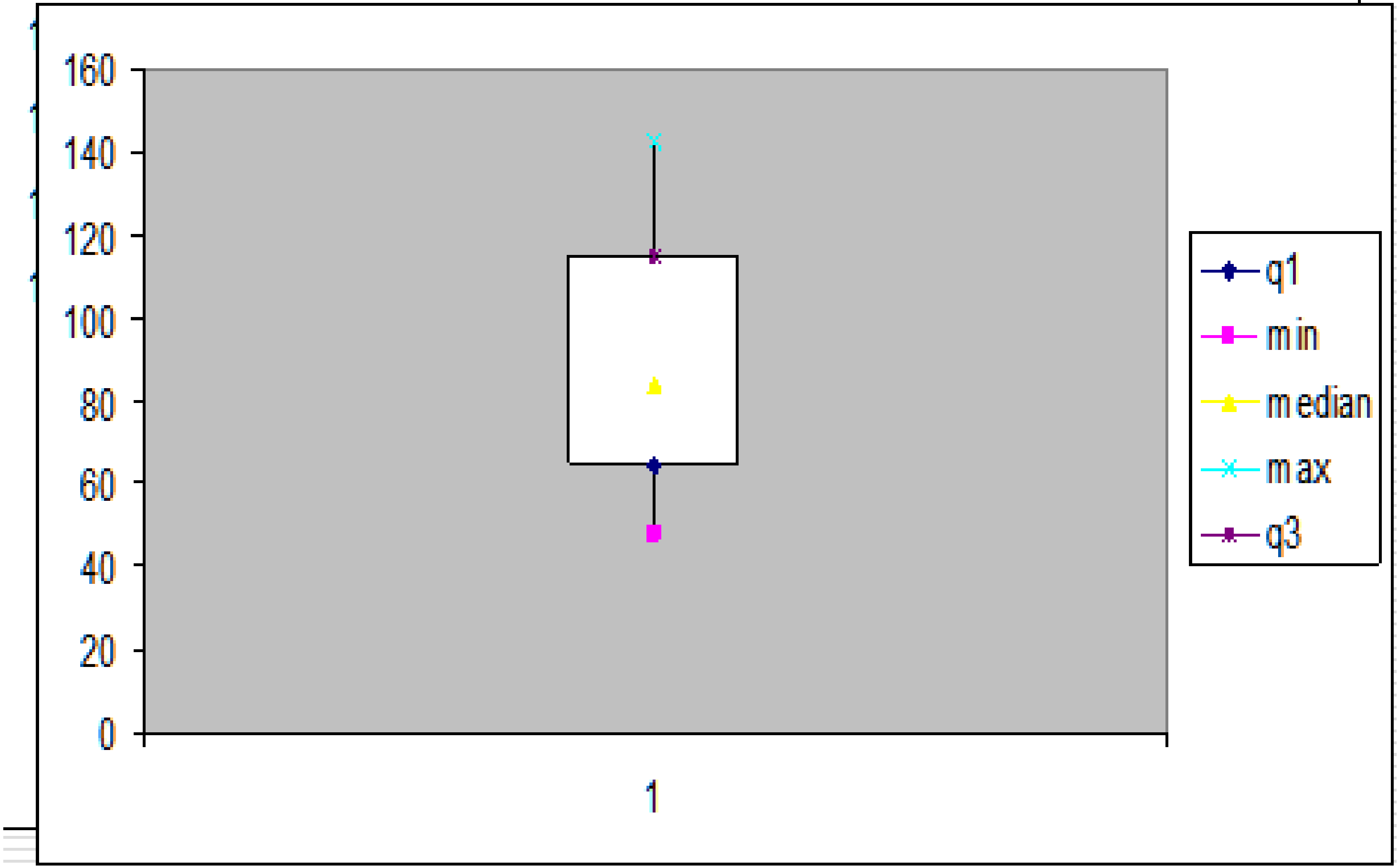
Box Plot: A box plot is a graph of the five number summary.

The central box spans the quartiles.

A line within the box marks the median.

Lines extending above and below the box mark the smallest and the largest observations (i.e., the range).

Outlying samples may be additionally plotted outside the range.


160
140
120
100
80
60
40
20
0
1
q1
min
median
max
q3

Distribution - of a variable tells us what values the variable takes and how often it takes these values.

- Unimodal - having a single peak
- Bimodal - having two distinct peaks
- Symmetric - left and right half are mirror images.

Measures of Central Tendency

Center measurement is a summary measure of the overall level of a dataset

- **Mean**
- **Median**
- **Mode**
- **Geometric mean**
- **Harmonic Mean**

Let $x_1, x_{2,} ... x_n$ are n observations of a variable x.

Then the mean of this variable,

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

For frequency data

$$\bar{x} = \frac{f_1 x_1 + f_2 x_2 + ... + f_n x_n}{\sum_{i=1}^{n} f_i} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}$$

Mean

Note that if the data is grouped data then each $\mathrm{x_i}$ is the class representative of that class

i.e., $\mathrm{x_i} = \frac{x_i^l + x_i^h}{2}$, where

x_i^l : lower bound for $\mathrm{i^{th}}$ class

x_i^h : upper bound for $\mathrm{i^{th}}$ class

Median: The middle value in an ordered sequence of observations.

Mode : The value that is observed most frequently.

The mode is undefined for sequences in which no observation is repeated.

The median is less sensitive to outliers (extreme scores) than the mean and thus a better measure than the mean for highly skewed distributions.

Mean – Mode = 3 (Mean – Median)

Median

Median for Individual series

In individual series, where data is given in the raw form, arrange the data in ascending or descending order.

- If the value of **N** is odd then simply the value of **(N+1)/2** th item is median for the data.
- If the value of **N** is even, then

  Median = [(N+1)/2 item + (N/2 + 1)th item]/2

Median

Median from Discrete Data

When the data follows a discrete set of values,

use $\left(\frac{N+1}{2}\right)^{th}$ item for finding the median.

First form a cumulative frequency distribution.

The median is that value which corresponds to

the cumulative frequency in which $\left(\frac{N+1}{2}\right)^{th}$ item lies.

Median

Grouped Data

Step 1: Construct the cumulative frequency distribution.

Step 2: Decide the class that contain the median.

Class Median is the first class with the value of cumulative frequency equal at least n/2.

Step 3: Find the median by using the following formula:

$$\text{Median} = L_m + \left(\frac{\frac{n}{2} - F}{f_m} \right) i$$

Where:

n = the **total frequency**

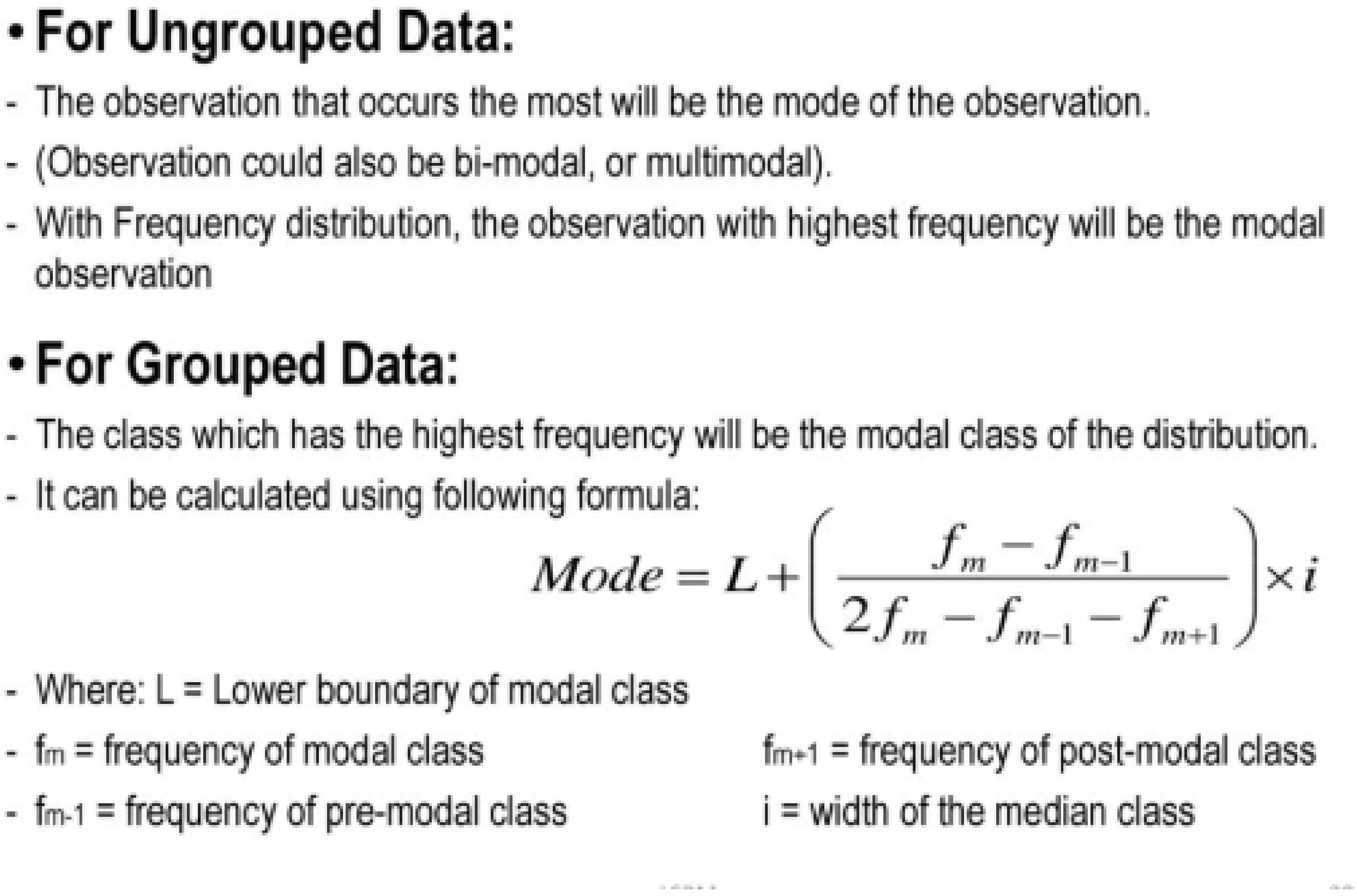
F = the **cumulative frequency** ***before*** class median

f_m = the **frequency** of the class median

i = the class width

L_m = the **lower boundary** of the class median

Mode



- **For Ungrouped Data:**

- The observation that occurs the most will be the mode of the observation.
- (Observation could also be bi-modal, or multimodal).
- With Frequency distribution, the observation with highest frequency will be the modal observation

- **For Grouped Data:**

- The class which has the highest frequency will be the modal class of the distribution.
- It can be calculated using following formula:

$$Mode = L + \left(\frac{f_m - f_{m-1}}{2f_m - f_{m-1} - f_{m+1}} \right) \times i$$

- Where: L = Lower boundary of modal class
- f_m = frequency of modal class
- f_{m-1} = frequency of pre-modal class

f_{m+1} = frequency of post-modal class

i = width of the median class

Shape of Data

measured by

- ❖ Skewness :

  Lack of symmetry

- ❖ Kurtosis

  Degree of peakedness

Moments about Mean

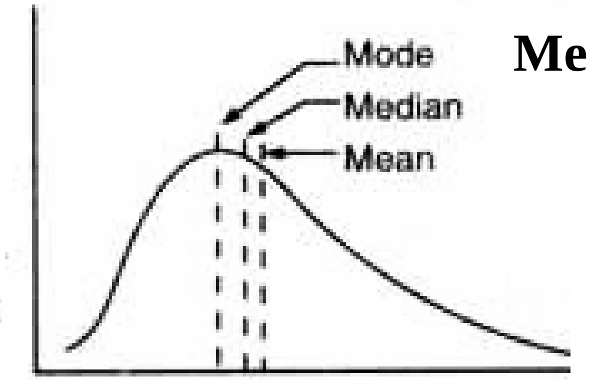
r^{th} moment μ_r about mean $\overline{x}$ is given by

$$\mu_r = \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - \overline{x})^r \text{, where}$$

$$N = \sum_{i=1}^{n} f_i = \text{total number of observations}$$

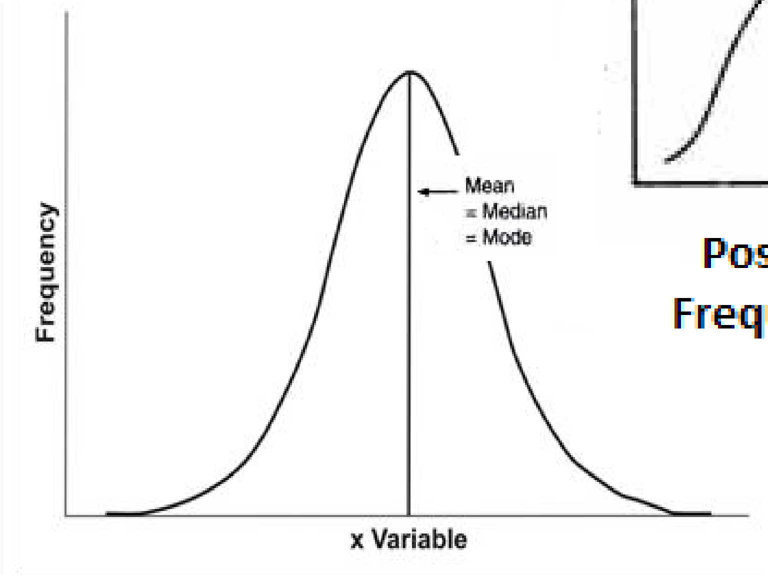
Skewness:: measures asymmetry of data

- ➢ Positive or right skewed: Longer right tail
- ➢ Negative or left skewed: Longer left tail

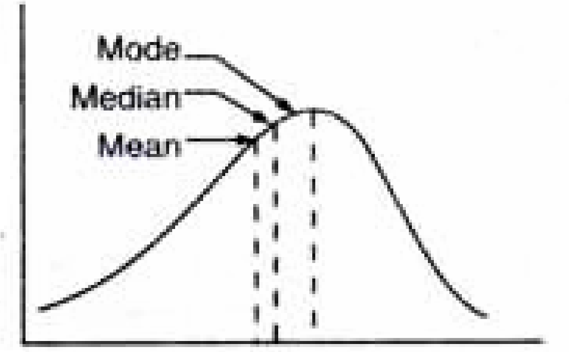


Positively Skewed
Frequency Distribution

Mean > Median > Mode



Mean = Median = Mode



Negatively Skewed
Frequency Distribution

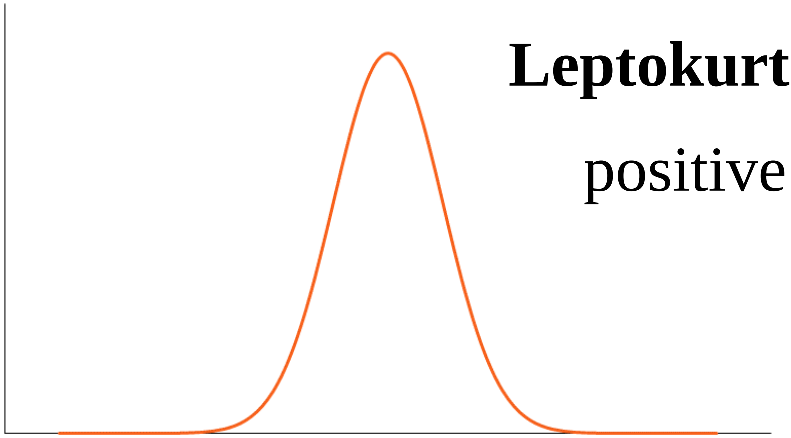
Mean < Median < Mode

Kurtosis measures peakeness of the distribution

The kurtosis of normal distribution is 0.

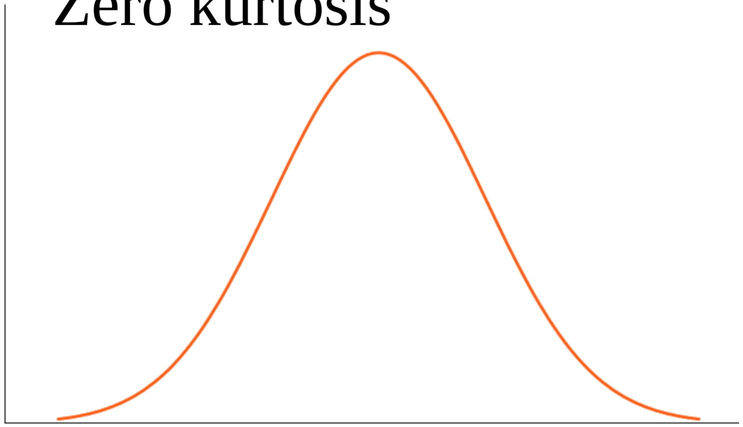
Leptokurtic

positive excess kurtosis



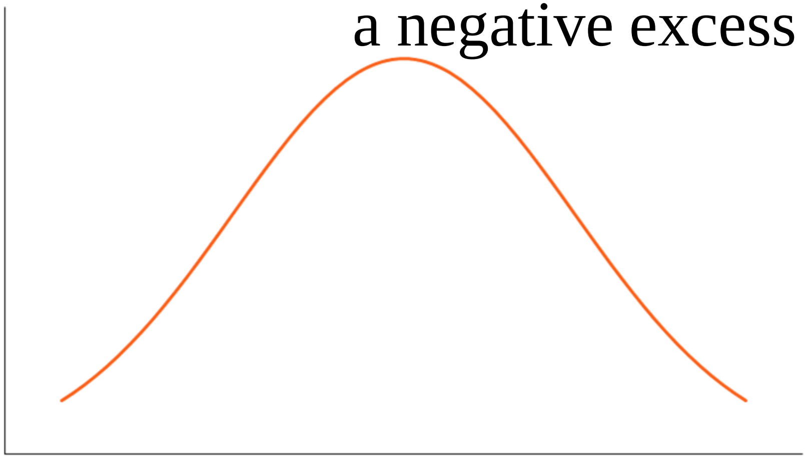
Mesokurtic

Zero kurtosis



Platykurtic

a negative excess kurtosis



Karl Pearson's Coefficient

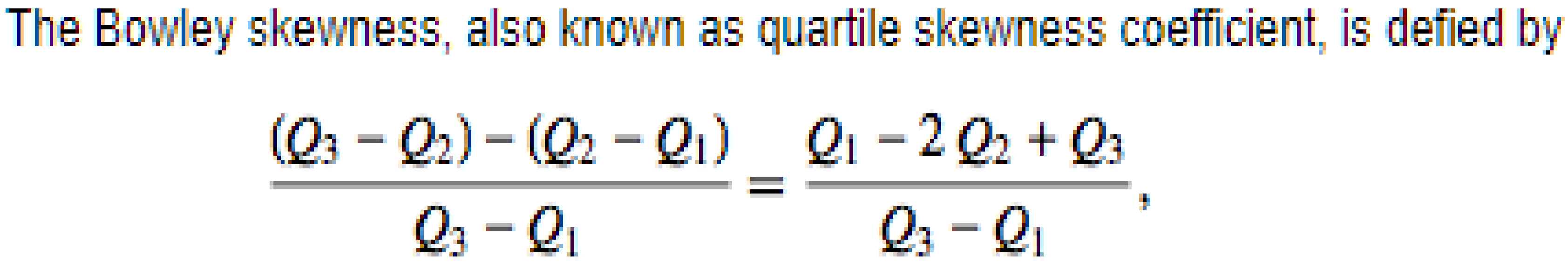
- Karl pearson's coefficient of skewness (Mode) is denoted by S_k, is given by,

$$S_k = \frac{Mean - Mode}{Standard\ deviation}$$

- Karl pearson's coefficient of skewness (Median) is denoted by Sk, is given by,

$$S_k = \frac{3(Mean - Median)}{Standard\ Deviation}$$

Bowley’s Coefficient of Skweness



The Bowley skewness, also known as quartile skewness coefficient, is defied by

$$\frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1} = \frac{Q_1 - 2\,Q_2 + Q_3}{Q_3 - Q_1},$$

Note that

Coefficient of Skewness $S_k > 0 \Rightarrow$

data is positively skewed.

Coefficient of Skewness $S_k < 0 \Rightarrow$

data is negatively skewed.

The coefficient of Kurtosis or Kurtosis is given by

$$\beta_2 = \frac{\mu_4}{\mu_2^2}, where$$

μ_4 : 4th moment about mean, μ_2 :second moment or variance.

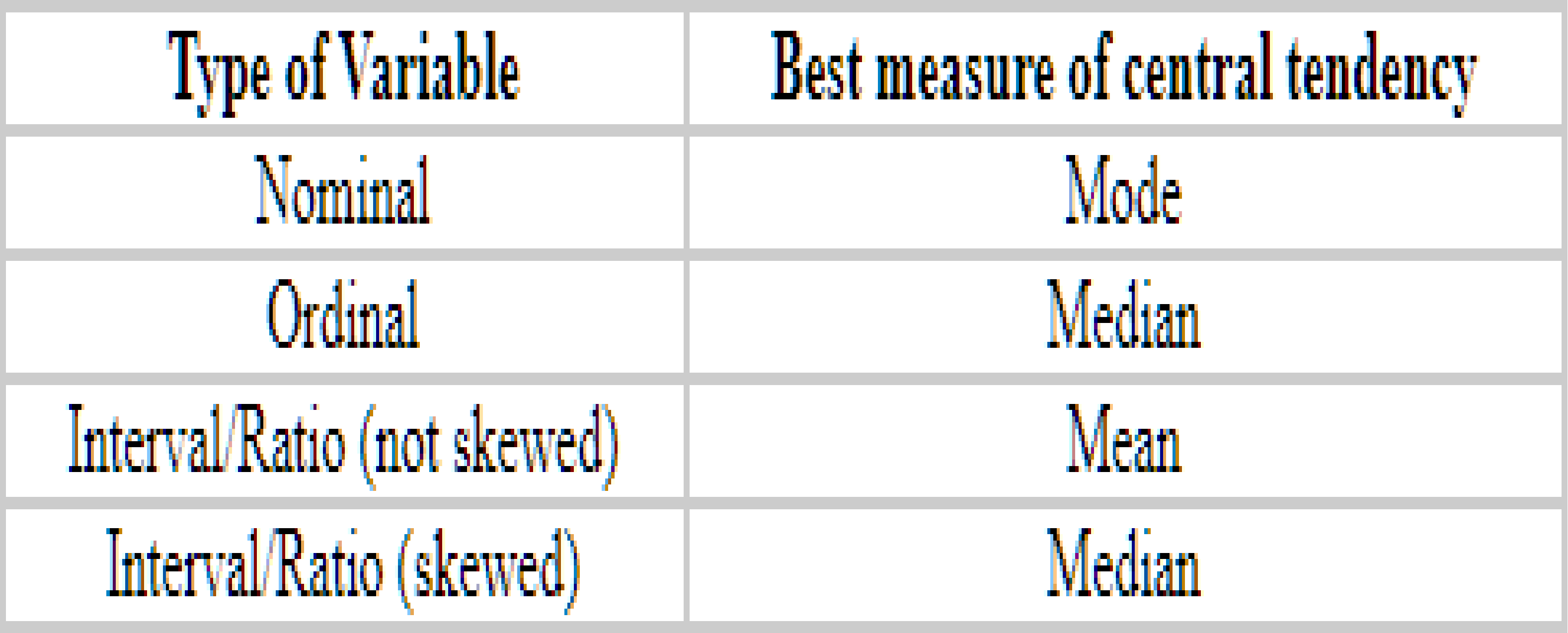
Note that :

1. If $\beta_2 = 3$, the data is normal or mesokurtic

2. If $\beta_2 > 3$, the data is peaked or leptokurtic

3. If $\beta_2 < 3$, the data is flat topped or platykurtic

Type of Variable Vs Measures of Central Tendency



Type of Variable	Best measure of central tendency
Nominal	Mode
Ordinal	Median
Interval/Ratio (not skewed)	Mean
Interval/Ratio (skewed)	Median

Methods of Variability Measurement

Variability (or dispersion) measures the amount of scatter in a dataset.

Commonly used methods: *range*, *variance*, *standard deviation*, *inter quartile range*, *coefficient of variation etc*.

Range: The difference between the largest and the smallest observations.

Variance:

Let $x_1, x_{2,} ...x_n$ are n observations of a variable x.

Then the variance of this variable,

$$Var(X) = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}{n}$$

For frequency data

$$Var(X) \quad = \quad \frac{\sum_{i=1}^{n} f_i \left(x_i - \overline{x}\right)^2}{\sum_{i=1}^{n} f_i}$$

Standard Deviation: Square root of the variance.

$$\sigma_X = +\sqrt{Var(X)}$$

Quartiles: Data can be divided into four regions that cover the total range of observed values.
Cut points for these regions are known as **quartiles**.

Quartiles of a data is the ((n+1)/4)qth observation of the data, where q is the desired quartile and n is the number of observations of data.

The first quartile (Q1) is the first 25% of the data.

The second quartile (Q2) is between the 25th and 50th percentage points in the data.

The upper bound of Q2 is the median.

The third quartile (Q3) is the 25% of the data lying between the median and the 75% cut point in the data.

Q1 is the median of the first half of the ordered observations and

Q3 is the median of the second half of the ordered observations.

Inter-quartile Range: Difference between Q3 and Q1.

Deciles: If data is ordered and divided into 10 parts, then cut points are called Deciles

Percentiles: If data is ordered and divided into 100 parts, then cut points are called Percentiles.

25th percentile is the Q1, 50th percentile is the Median (Q2) and the 75th percentile of the data is Q3.

Coefficient of Variation: The standard deviation of data divided by it’s mean. It is usually expressed in percent.

$$Coefficient\ of\ \ Variation = \frac{\sigma_X}{\overline{x}} \times 100$$

Softwares to perform statistical analysis and visualization of data.

SAS (System for Statistical Analysis), S-plus, R, Matlab, Minitab, BMDP, Stata, SPSS, StatXact, Statistica, LISREL, JMP, GLIM, HIL, MS Excel etc.

Some useful websites for more information of statistical softwares

http://www.galaxy.gmu.edu/papers/astr1.html

http://ourworld.compuserve.com/homepages/Rainer_Wuerlaender/statsoft.htm#archiv

http://www.R-project.org