Welcome

Descriptive Statistics

Measures of location and dispersion

- **Md. Fazlul Karim Patwary**
- Associate Professor, IIT, JU

What is Statistics?

– …a set of procedures and rules…for reducing large masses of data to manageable proportions and for allowing us to draw conclusions from those data

What is Statistics?

- Statistics is the science of collection, organization, presentation, analysis and interpretation of data.
- A descriptive measure computed from the data of a sample is called a statistics.
- A statistics is a summary value of a sample (i.e. sample mean, median, mode, standard deviation etc.)

Q. What is Bio-statistics? Write the use of biostatistics in biological fields.

Types of Statistics

- Descriptive Statistics:
  - Descriptive statistics are used to describe the basic features of the data in a study.
  - Provide simple summaries about the sample and the measures.
  - With simple graphics analysis, it forms the basis of quantitative analysis of data.
- Inferential Statistics
  - Inference from sample to population
  - Inference from statistic to parameter
  - Factors influencing the accuracy of a sample's ability to represent a population:
    - Size
    - Randomness

    Q. What is inferential Statistics? Write the importance of this.

Attribute and Variables

Variable: Which can take any values during the experiment or trial.

i.e. if x represent age of student then recording of any one student’s age must be different. That is value of x is different.

Attribute: A characteristic of an object or entity i.e. colour of eyes of the tourists.

Types of Variables

Discrete Variable: Which can take only discrete values during the experiment or trial.

i.e. if x represent number of goals in football tournament then recording of goals of a tournament must be integer. Here, x is a discrete variable.

Continuous Variable: Which can take any value within a possible range during the experiment or trial.

i.e. if x represent study time per day of students then x may be between 0-24 hours.

Frequency distribution

Table 2.2 Net Weight in Ounces of Fruit

19.7	19.9	20.2	19.9	20.0	20.6	19.3	20.4	19.9	20.3
20.1	19.5	20.9	20.3	20.8	19.9	20.0	20.6	19.9	19.8

Table 2.3 Frequency Distribution of Weights

Weight, oz	Class Midpoint	Absolute Frequency	Relative Frequency	Cumulative Frequency
19.2–19.4	19.3	1	0.05	1
19.5–19.7	19.6	2	0.10	3
19.8–20.0	19.9	8	0.40	11
20.1–20.3	20.2	4	0.20	15
20.4–20.6	20.5	3	0.15	18
20.7–20.9	20.8	2	0.10	20
		20	1.00	

Frequency distribution: SPSS Data

*Untitled1 [DataSet0] - SPSS Data Editor

File Edit View Data Transform Analyze Gr

1 : age 39

	age	age_a
1	39	4.00
2	20	1.00
3	26	2.00
4	23	1.00
5	27	2.00
6	29	2.00
7	30	2.00
8	32	3.00
9	31	3.00
10	34	3.00
11	35	3.00
12	36	4.00
13	23	1.00
14	23	1.00
15	34	3.00
16	40	4.00
17	37	4.00
18	38	4.00
19	29	2.00

Frequency distribution: SPSS Command

```
recode age
    (20 thru 25=1)
    (26 thru 30=2)
    (31 thru 35=3)
    (36 thru 40=4)
into age_a.

add value label age_a
    1 '20 - 25'
    2 '26 - 30'
    3 '31 - 35'
    4 '36 - 40'.
execute.


FREQUENCIES VARIABLES=age_a
  /ORDER=ANALYSIS.
```

Frequency distribution: SPSS Command

age_a

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	20 - 25	6	18.2	18.2	18.2
	26 - 30	6	18.2	18.2	36.4
	31 - 35	8	24.2	24.2	60.6
	36 - 40	13	39.4	39.4	100.0
	Total	33	100.0	100.0	

- Q. Comment on/describe the above results

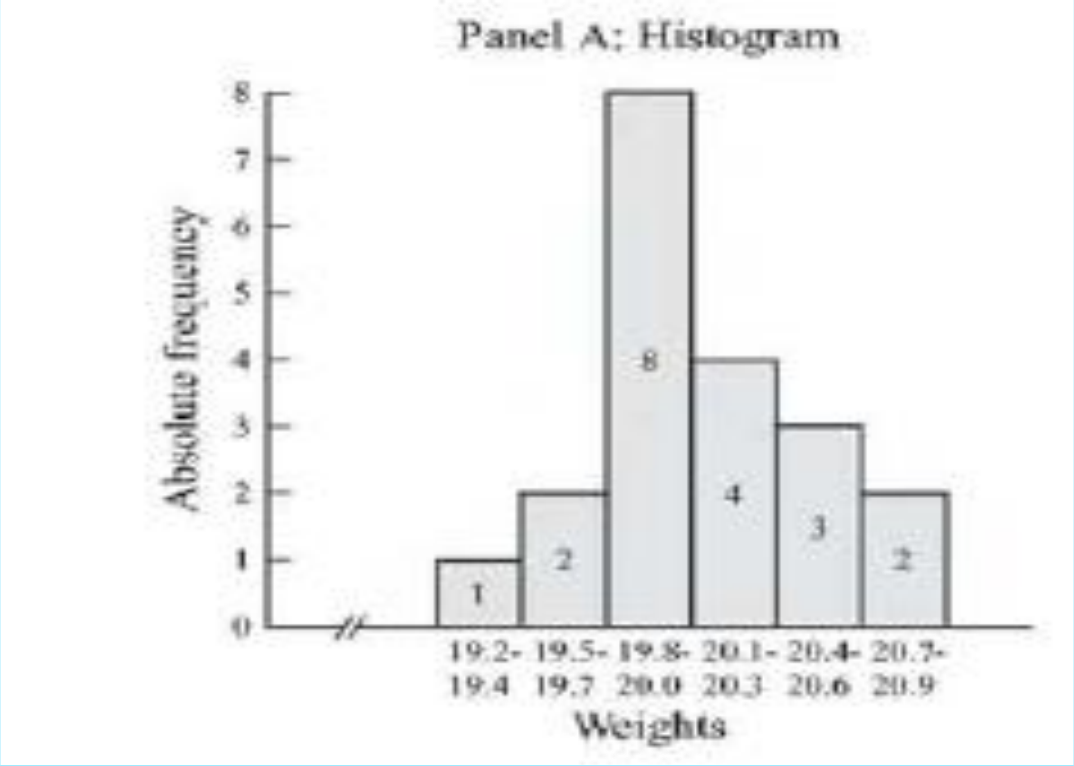
Frequency distribution

- Presentation of data into groups or classes
- Shows the number of observations in each class.

- Relative frequency distribution - % of frequency
- Cumulative frequency distribution – cumulative % of frequency
- Histogram - Graphical presentation

Frequency distribution

- What are the difference between bar Chart and Histogram
- When do we use them?



Measures of Location or Central Tendency

- A measure of central tendency (also referred to as **measures of centre** or **central location**) is a summary measure that attempts to describe a whole set of data with a single value that represents the middle or centre of its distribution.

- The term "**measures of central tendency**" refers to finding the **mean, median and mode**.

Most important measures of central tendency are:

1. Mean,
2. Median and
3. Mode.

Central Tendency: Mean

- Averages
  - Mode: most frequently occurring value in a distribution (any scale, most unstable)
  - Median: midpoint in the distribution below which half of the cases reside (ordinal and above)
  - Mean: arithmetic average- the sum of all values in a distribution divided by the number of cases (interval or ratio)

Central Tendency: Mean

- Let $x_1, x_2, x_3, \ldots, x_n$ be the realised values of a random variable **X**, from a sample of size **n.** The **sample arithmetic mean** is defined as:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Central Tendency: Mean

- Mean is half the sum of a set of values:
- Scores: 5, 6, 7, 10, 12, 15
- Sum: 55
- Number of scores: 6
- Computation of Mean: 55/6= **9.17**

Central Tendency: Mean (group Data)

- Mean of a group data can be calculated by the formula:
- Mean = $\bar{x} = \frac{\sum fx}{n}$

Central Tendency: Mean (group Data)

Age	Frequency , f
0-4	10
5-9	12
10-14	6
15-19	2
Total	30

This data is grouped into 4 class intervals of width 4.
The data is discrete.

Central Tendency: Mean (group Data)

Age	Frequency (f)	Midpoint (x)	f * x	Cumulative Frequency (cf)
0-4	10	2	20	10
5-9	12	7	84	22 (=10+12)
10-14	6	12	72	28 (=22+6)
15-19	2	17	34	30 (=28+2)
Total	30	38	210	

Central Tendency: Mean (group Data)

- Mean = $\bar{x} = \frac{\sum fx}{n} = \frac{20+84+72+34}{10+12+6+2} = \frac{210}{30} = 7$

```
FREQUENCIES VARIABLES=age
  /FORMAT=NOTABLE
  /STATISTICS=MEAN median mode
  /ORDER=ANALYSIS.
```

SPSS Output

Statistics

age

N	Valid	30
	Missing	0
Mean	31.30	

Central Tendency: Median

- If the sample data are arranged in increasing order, the **median** is
  - (i) the <u>middle</u> value if *n* is an odd number, or
  - (ii) <u>midway</u> between the two middle values if *n* is an even number
- The **mode** is the most commonly occurring value.

Central Tendency: Median

- Example (11 test scores)

61, 61, 72, 77, 80, 81, 82, 85, 89, 90, 92

The median is 81 (half of the scores fall above 81, and half below)

Central Tendency: Median

- Example (6 scores)

3, 3, 7, 10, 12, 15

Even number of scores= Median is half-way between these scores

Sum the middle scores (7+10=17) and divide by 2

17/2= **8.5**

Central Tendency: Median

- Insensitive to extremes

3, 3, 7, 10, 12, 15, 200

Central Tendency: Median (group data)

- Median is the positional average.
- Median is affected by the extreme value.
- Mean of group data can be calculate by:

- Median = $L_1 + \dfrac{\frac{n}{2} - CF}{fm} \times i$

Here,

L_1 = Lower limit of the median class
CF = Cumulative frequency prior to median group
fm = frequency of the median group
i = Class interval of the median group
n = total frequency

How to find Median Class:

1. Find the value of n/2
2. Check from cumulative frequency, where does the value of n/2 fall. The class that has CF lie is the Median class

Class Interval: The difference between upper and lower limit of class is known as class interval

Central Tendency: Median (group data)

Age	Frequency, f	(Midpoint Age), m	fxm	Cumulative Frequency
0-4	10	2	20	10
5-9	12	7	84	22 (=10+12)
10-14	6	12	72	28 (=22+6)
15-19	2	17	34	30 (=28+2)
Total	30	38	210	

Median Class 5 - 9

Central Tendency: Median (group data)

- Median can be calculated from a group data by:

- Median = $L_1 + \dfrac{\frac{n}{2} - CF}{fm} \times i$

$$= 5 + \frac{\frac{30}{2} - 10}{12} \times 4 = 5 + \frac{15 - 10}{12} \times 4 = 5 + \frac{5}{12} \times 4 = 5 + 0.416 \times 4 = 5 + 1.66 = 6.66$$

Central Tendency: Mode

- Mode is the most frequently occurring value in a set.
- Best used for nominal data.
- We cannot find an exact value for the mode, and therefore give the **modal class**. The modal class is 5-9 for group data.

Central Tendency: Mode (Group Data)

- Mode can be calculated from a group data by:

- Mode = $L_1 + \dfrac{\Delta_1}{\Delta_1 + \Delta_2} \times i$

Here,

L_1 = Lower limit of the modal class (modal **class** where highest number of people lie)

Δ_1 = Difference between the frequency of the modal class and its preceding class

Δ_2 = Difference between the frequency of the modal class and its following class

i = Class interval of the modal class

Central Tendency: Mode (Group Data)

- Mode is the most frequently occurring value in

Age	Frequency, f	(Midpoint Age), m	fxm	Cumulative Frequency
0-4	10	2	20	10
5-9	12	7	84	22 (=10+12)
10-14	6	12	72	28 (=22+6)
15-19	2	17	34	30 (=28+2)
Total	30	38	210	

Modal Class 5 - 9

The modal class is simply the class interval of **highest frequency.**

Central Tendency: Mode (Group Data)

- Mode can be calculated from a group data by:
- Mode = $L_1 + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i$

$$= 5 + \frac{12-10}{(12-10)+(12-6)} \times 4 = 5 + \frac{2}{2+6} \times 4 = 5 + \frac{2}{8} \times 4$$

$$= 5 + 0.25 \times 4 = 5 + 1 = 6$$

Variability

- Variability is the differences among scores- shows how subjects vary:
  - Dispersion: extent of scatter around the "average"
  - Range: highest and lowest scores in a distribution
  - Variance and standard deviation: spread of scores in a distribution. The greater the scatter, the larger the variance
- Interval or ration level data
- **Standard deviation:** how much subjects differ from the mean of their group

Measures of dispersion

A measure of statistical dispersion is a nonnegative real number that is zero if all the data are the same and increases as the data become more diverse.

- Standard deviation
- Interquartile range (IQR) or Interdecile range
- Range
- Mean difference
- Median absolute deviation (MAD)
- Average absolute deviation (or average deviation)
- Distance standard deviation

Relative and absolute measures of dispersion

- **Absolute Measure of dispersion**: Variation are calculated from the mean

i.e. Standard deviation.

- **Relative measure of dispersion**: These are the position of certain variable as compared with the other variables.

i.e. percentiles, quartiles or the z-score.

Measures of dispersion

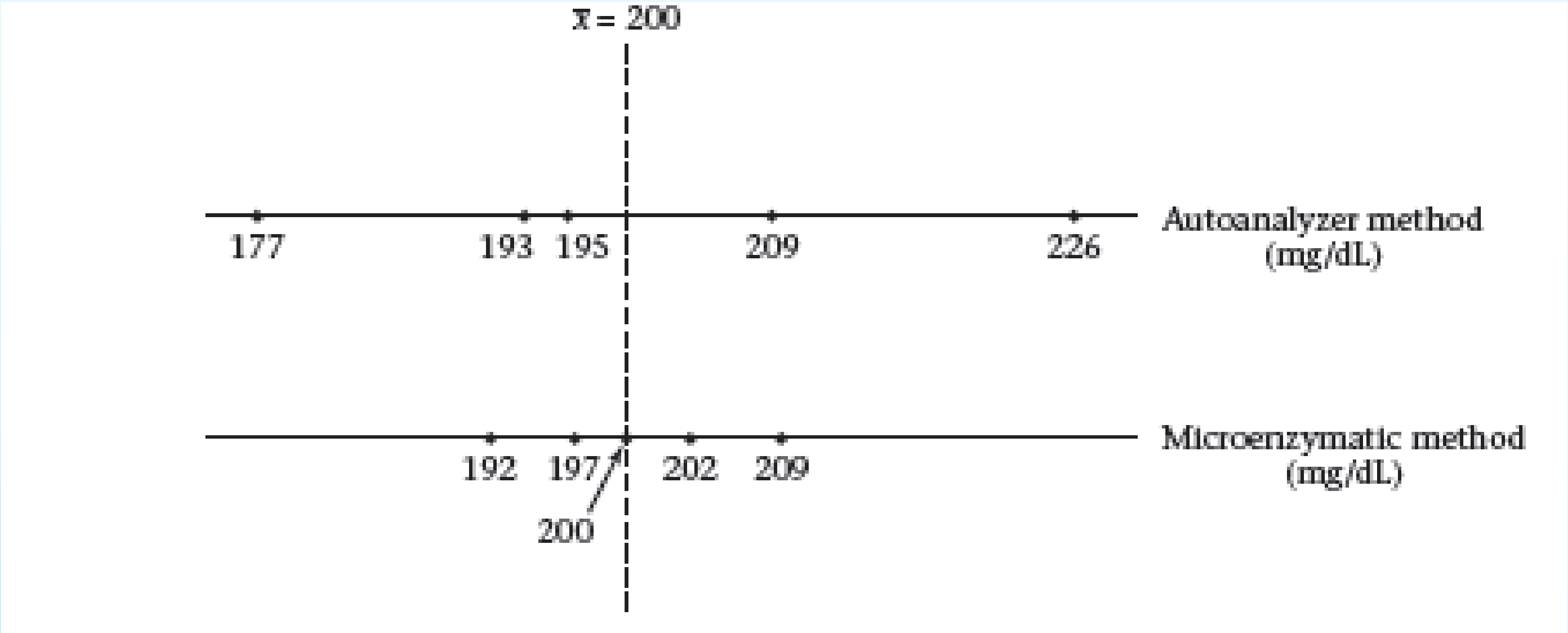
- **Range**: The range is the difference between the largest and smallest observations in a sample.

Sample of birthweights (g) of live-born infants born at a private hospital in San Diego, California, during a 1-week period

i	x_i	i	x_i	i	x_i	i	x_i
1	3265	6	3323	11	2581	16	2759
2	3260	7	3649	12	2841	17	3248
3	3245	8	3200	13	3609	18	3314
4	3484	9	3031	14	2838	19	3101
5	4146	10	2069	15	3541	20	2834

Range= 4146-2069=2077 g

Measures of dispersion



Autoanalyzer method: Range= 226 - 177= 49 mg/dl

Microenzymatic method: Range= 209 - 192= 17 mg/dl

Measures of dispersion

Quantile or percentile:

The pth percentile is defined by

1. The (k + 1)th largest sample point if np/100 is not an integer (where k is the largest integer less than np/100)
2. The average of the (np/100)th and (np/100 + 1)th largest observations if np/100 is an integer.

Measures of dispersion: Quantile or percentile

Sample of birthweights (g) of live-born infants born at a private hospital in San Diego, California, during a 1-week period

i	x_i	i	x_i	i	x_i	i	x_i
1	3265	6	3323	11	2581	16	2759
2	3260	7	3649	12	2841	17	3248
3	3245	8	3200	13	3609	18	3314
4	3484	9	3031	14	2838	19	3101
5	4146	10	2069	15	3541	20	2834

- Compute the 10th and 90th percentiles for the birth weight data.

20 × .1 = 2 and 20 × .9 = 18 are integers, the 10th and 90th percentiles are:

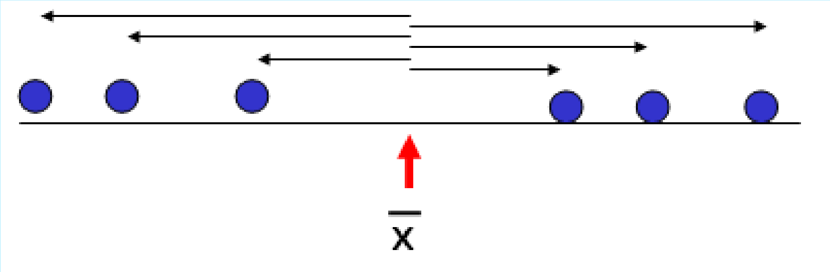
10th percentile: average of the second and third largest values
= (2581 + 2759)/2 = 2670 g

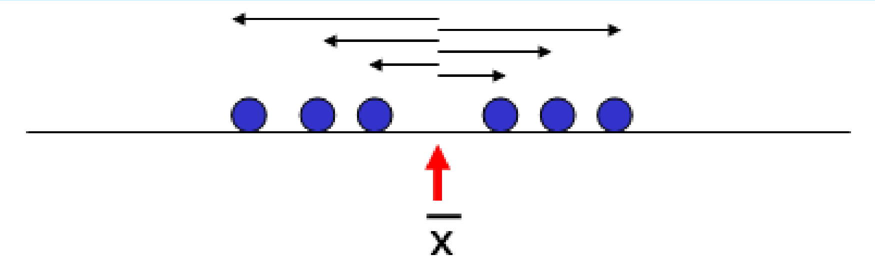
90th percentile: average of the 18th and 19th largest values
= (3609 + 3649)/2 = 3629 g

Standard Deviation

- The **sample variance, s^2**, is the arithmetic mean of the squared deviations from the sample mean:

$$s^2 = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1}$$





Standard Deviation

- The **sample standard deviation, s,** is the square-root of the variance

$$s = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1}}$$

- **s** has the advantage of being in the same units as the original variable **x**

Standard Deviation

Data	Deviation	Deviation^2
151	13.86	192.02
124	-13.14	172.73
132	-5.14	26.45
170	32.86	1079.59
146	8.86	78.45
124	-13.14	172.73
113	-24.14	582.88
Sum = 960.0	Sum = 0.00	Sum = 2304.86

Standard Deviation

$$\sum_{i=1}^{7}\left(x_i - \bar{x}\right)^2 = 2304.86$$

Therefore, $$s = \sqrt{\frac{2304.86}{7-1}}$$

$$= 19.6$$

Standard Deviation

- Measures how much subjects differ from the mean of their group
- The more spread out the subjects are around the mean, the larger the standard deviation
- Sensitive to extremes or “outliers”

Coefficient of Variation

- The **coefficient of variation** (CV) or **relative standard deviation** (RSD) is the sample standard deviation expressed as a percentage of the mean, i.e.

$$CV = \left(\frac{s}{\bar{x}} \right) \times 100\%$$

- The CV is not affected by multiplicative changes in scale
- Consequently, a useful way of comparing the dispersion of variables measured on different scales

Coefficient of Variation

The CV of the blood pressure data is:

$$CV = 100 \times \left(\frac{19.6}{137.1} \right) \%$$

$$= 14.3\%$$

i.e., the standard deviation is 14.3% as large as the mean.

Inter-quartile range

- The Median divides a distribution into two halves.

- The **first** and **third** quartiles (denoted $\mathbf{Q_1}$ and $\mathbf{Q_3}$) are defined as follows:
  - 25% of the data lie below Q_1 (and 75% is above Q_1),
  - 25% of the data lie above Q_3 (and 75% is below Q_3)

- The **inter-quartile range (IQR)** is the difference between the first and third quartiles, i.e.
  $\mathbf{IQR = Q_3 - Q_1}$

Inter-quartile range

The ordered blood pressure data is:

113 124 124 132 146 151 170

Q_1 ↑ Q_3 ↑

Inter Quartile Range (IQR) is 151-124 = 27

Box-Plots

- A box-plot is a visual description of the distribution based on
  - Minimum
  - Q1
  - Median
  - Q3
  - Maximum
- Useful for comparing large sets of data

Box-Plots

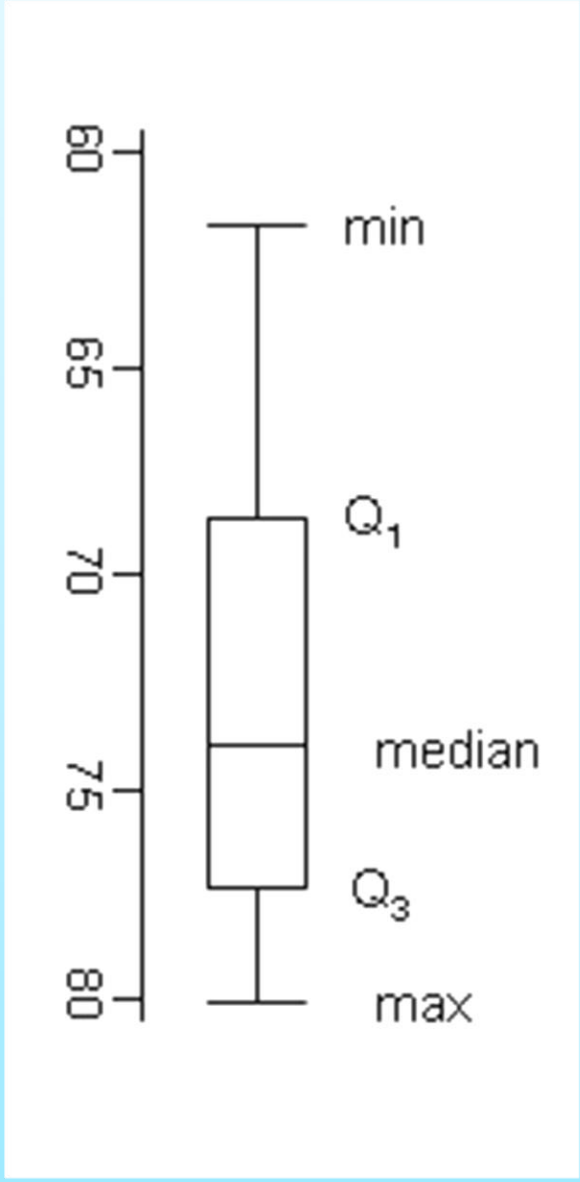
The pulse rates of 12 individuals arranged in increasing order are:

62, 64, 68, 70, 70, 74, 74, 76, 76, 78, 78, 80

$Q_1 = (68+70) \div 2 = 69$, $Q_3 = (76+78) \div 2 = 77$

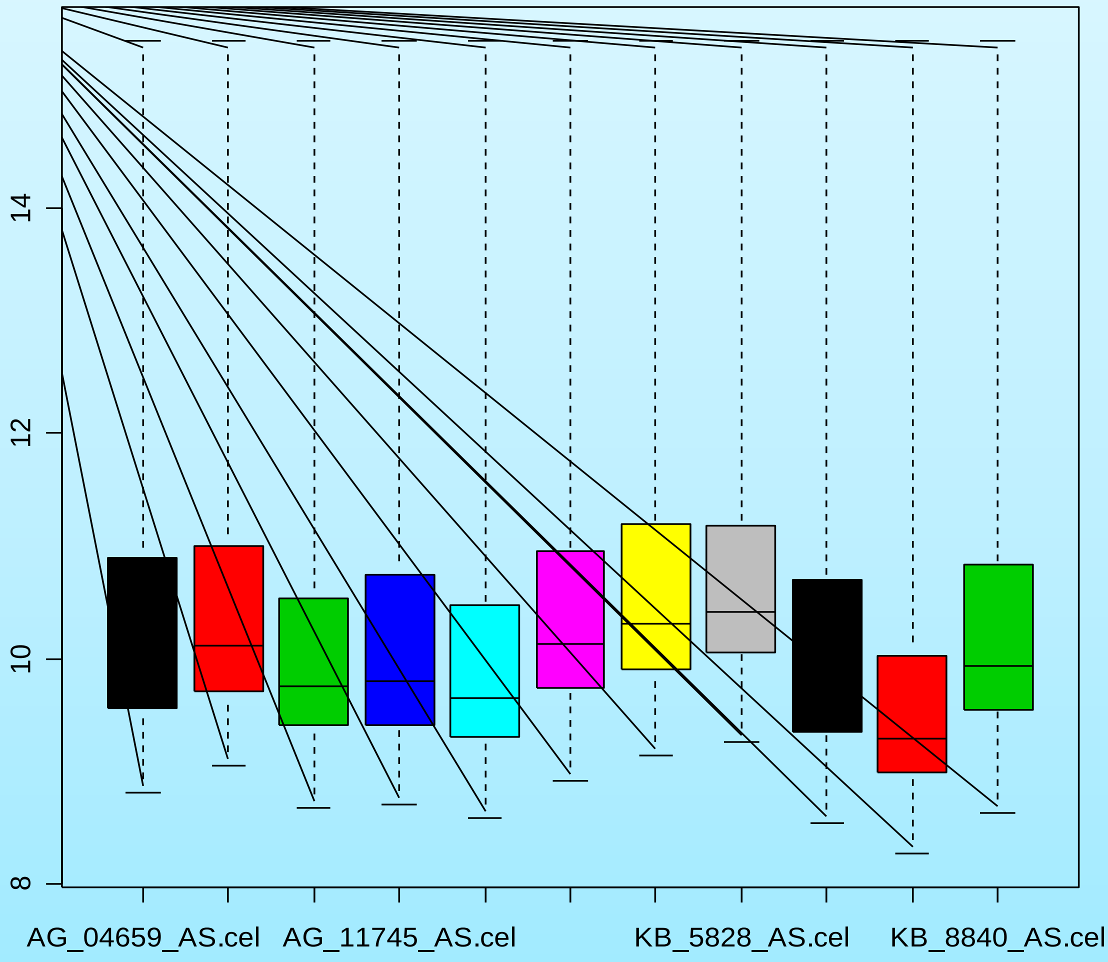
$IQR = (77 - 69) = 8$

Box-Plots



Box-Plots

Box-plots of intensities from 11 gene expression arrays



Outliers

- An **outlier** is an observation which does not appear to belong with the other data
- Outliers can arise because of a measurement or recording error or because of equipment failure during an experiment, etc.
- An outlier might be indicative of a sub-population, e.g. an abnormally low or high value in a medical test could indicate presence of an illness in the patient.

Outlier Boxplot

- Re-define the upper and lower limits of the boxplots (the whisker lines) as:

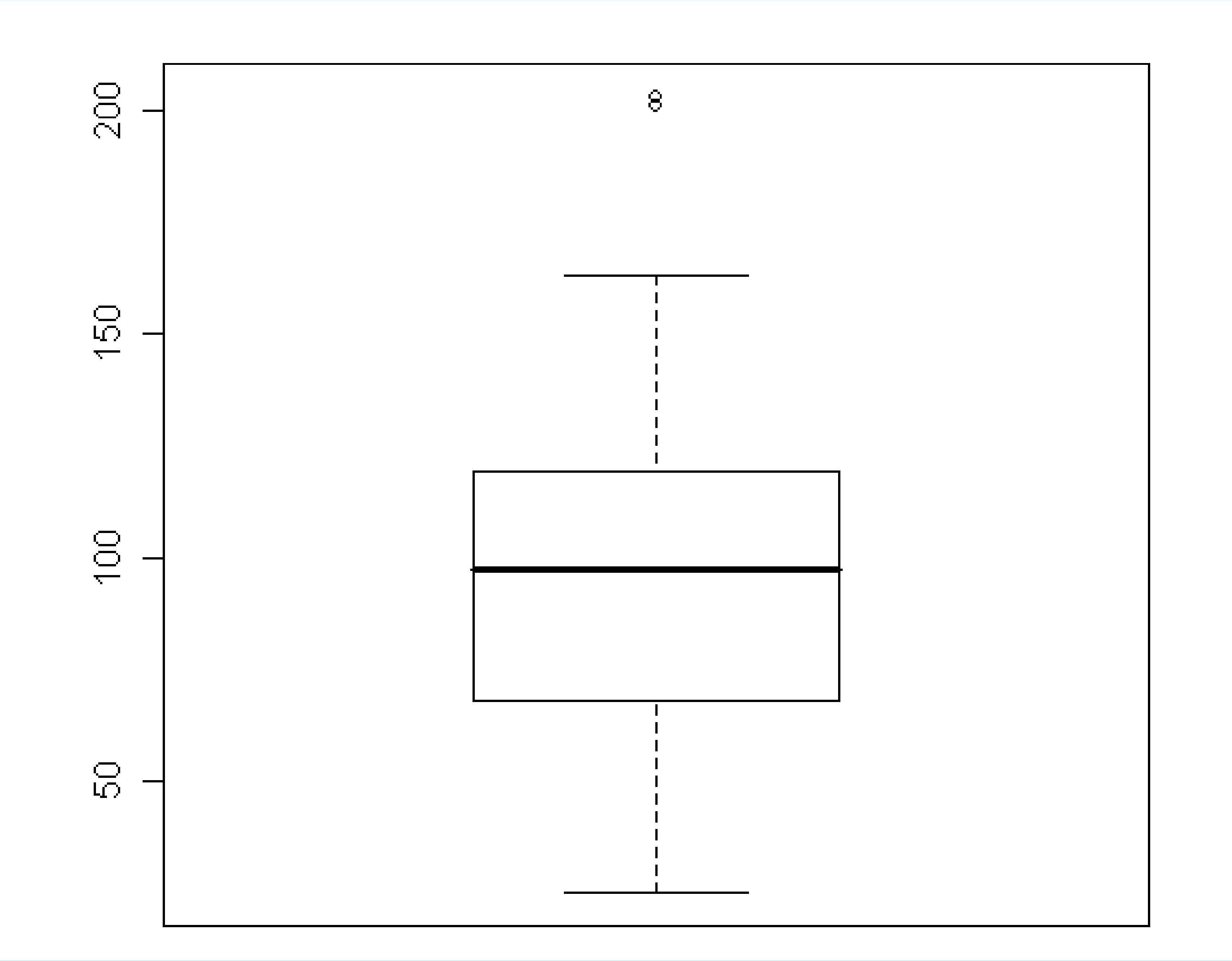
  Lower limit = $Q_1 - 1.5 \times IQR$, and

  Upper limit = $Q_3 + 1.5 \times IQR$

- Note that the lines may not go as far as these limits

- If a data point is < lower limit or > upper limit, the data point is considered to be an outlier.

Box-Plots



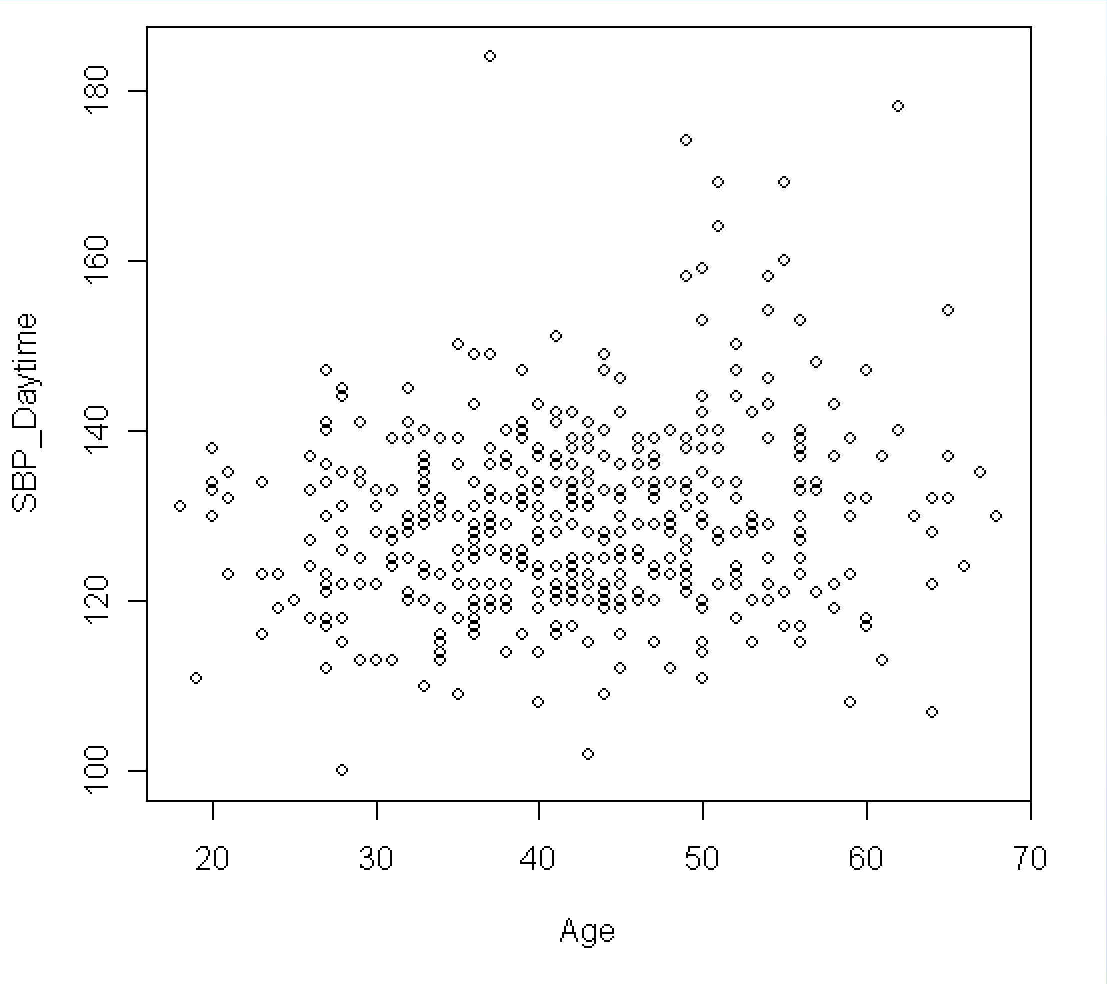
Scatter-plot

- Displays the relationship between two continuous variables

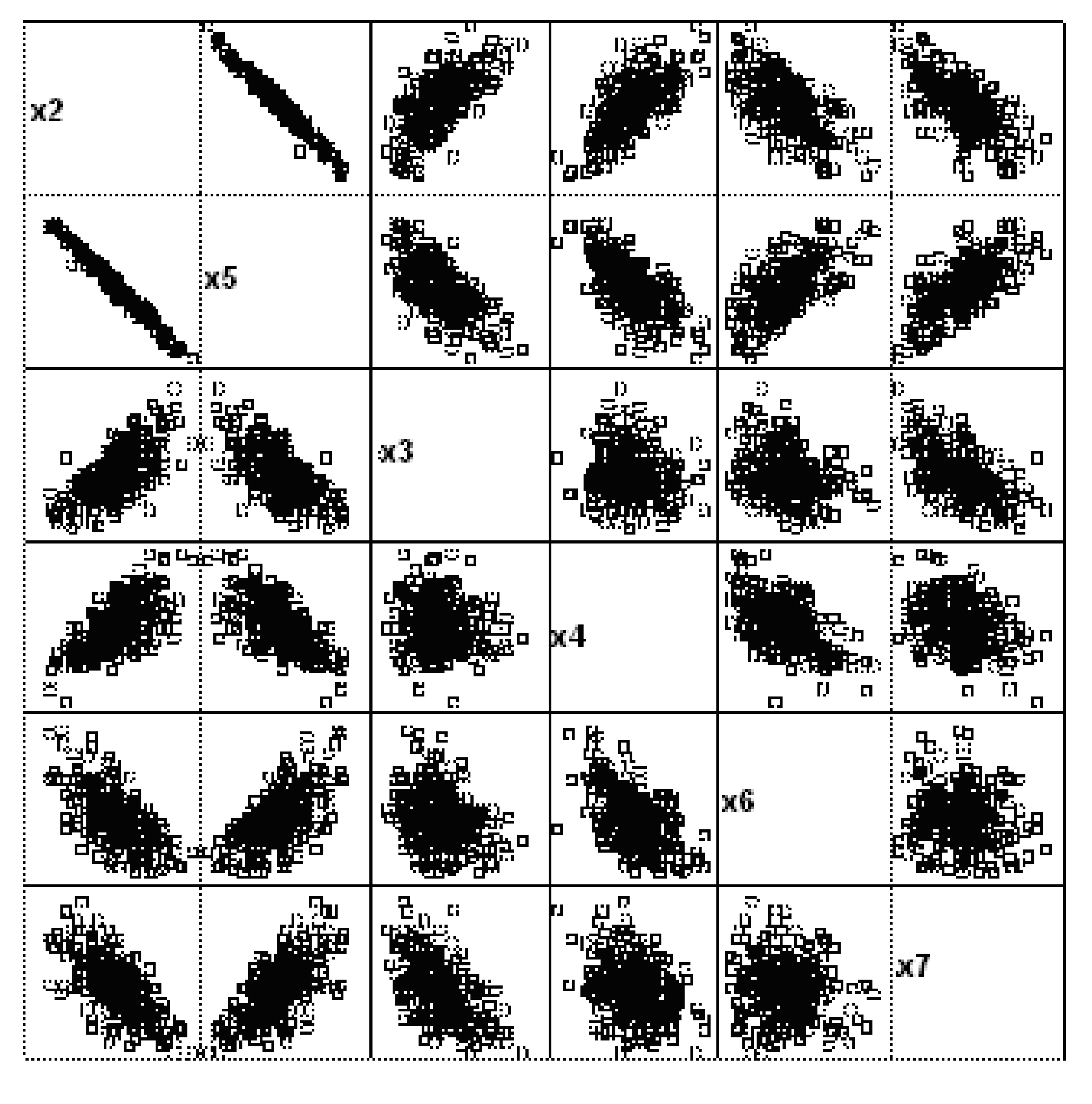
- Useful in the early stage of analysis when exploring data and determining is a linear regression analysis is appropriate

- May show outliers in your data

- Age versus Systolic Blood Pressure in a Clinical Trial



Scatter-plot matrix (multiple pair-wise plots)



Properties of Standard deviation

- When analyzing normally distributed data, standard deviation can be used in conjunction with the mean in order to calculate data intervals.

If x =mean, S=standard deviation and X=a value in the data set, then

1. about 68% of the data lie in the interval: $x-S<X<x+S$.
2. About 95% of the data lie in the interval: $x-2S<X<x+2S$.
3. About 99% of the data lie in the interval: $x-3S<X<x+3S$

Dispersion

- variability or spread in the data
- Most important measures of dispersion are :

1. Average deviation,
2. Variance, and
3. Standard deviation
4. Quartile coefficient dispersion
5. Inter-quartile range $IQR = Q_3 - Q_1$
6. Coefficient of Variation; CV=SD/Mean * 100

Shape of Frequency distribution

Gives the idea about:

- symmetry or lack of it (skewness) and
- Peakedness (ktirtosis]

Measures of Skewness:

1. Symmetrical
2. Positively skewed
3. Negatively skewed

Shape of Frequency distribution

Measures of Skewness:

Pearsons:

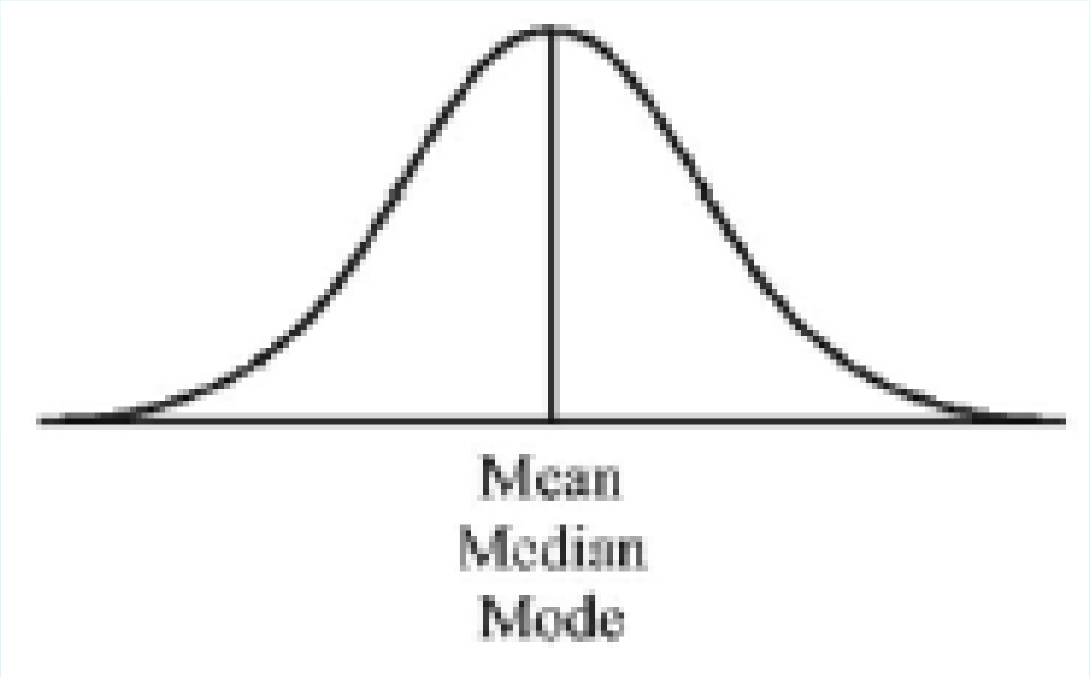
$$Sk = \frac{3(\mu - Median)}{\sigma}$$

$$Sk = \frac{1}{n}\frac{\sum(x_i - \mu)^3}{\sigma^3}$$

1. Symmetrical (SK=0)
2. Positively skewed (SK>0)
3. Negatively skewed (SK<0)

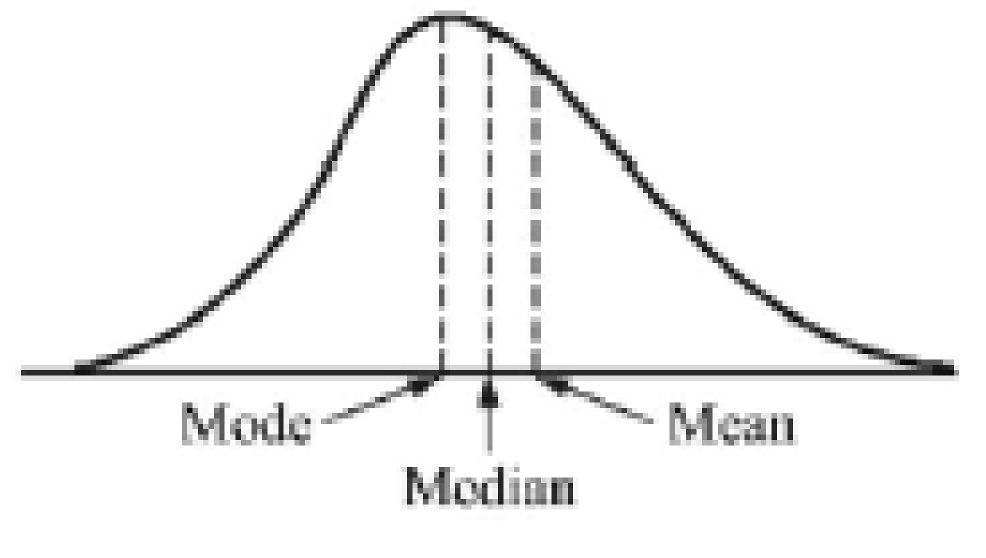
Shape of Frequency distribution

Skewness: A distribution is symmetrical if mean=median=mode



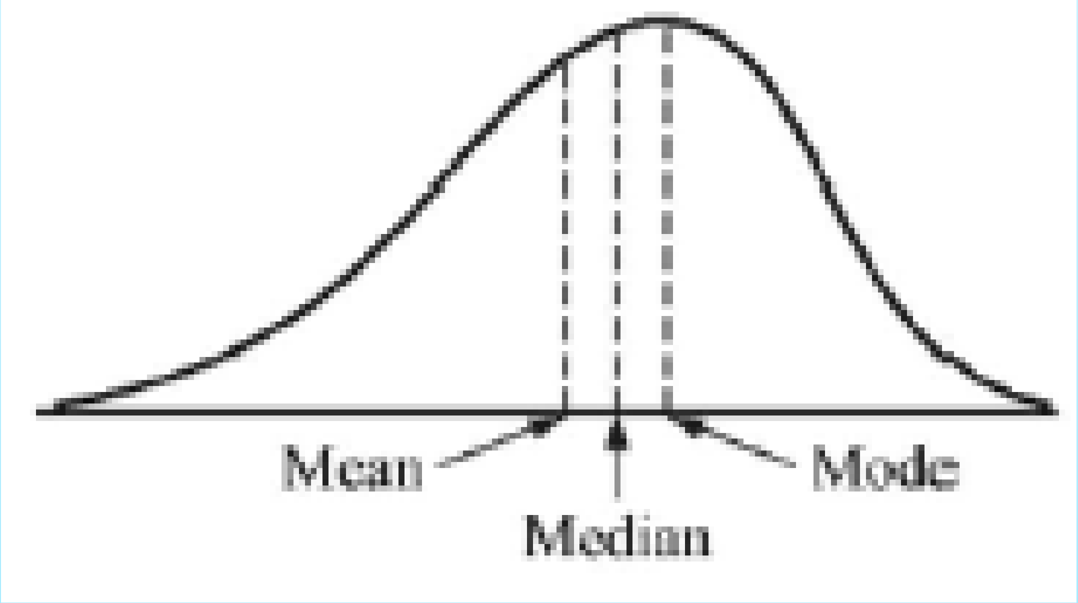
Shape of Frequency distribution

Skewness: A distribution is positively skewed if mean>median>mode



Shape of Frequency distribution

Skewness: A distribution is negatively skewed if mean<median<mode



Shape of Frequency distribution

- Pickness is measured by the term Kurtosis.
- Kurtosis is a measure of whether the data are peaked or flat relative to a normal distribution.

$$kurtosis = \frac{1}{n}\frac{\sum(x_i - \mu)^4}{\sigma^4}$$

Leptokurtic (kurtosis<3 or k=kurtosis-3<0)

Platykurtic (kurtosis>3 or k=kurtosis-3>0)

Mesokurtic (kurtosis=3 or k=kurtosis-3=0)

Example of Kurtosis