Descriptive Statistics: Numerical Summary Measures

- Single numbers which quantify the characteristics of a distribution of values
  - ✓ Measures of central tendency (location)
  - ✓ Measures of dispersion

- A frequency distribution is a general picture of the distribution of a variable

- But, can’t indicate the average value and the spread of the values

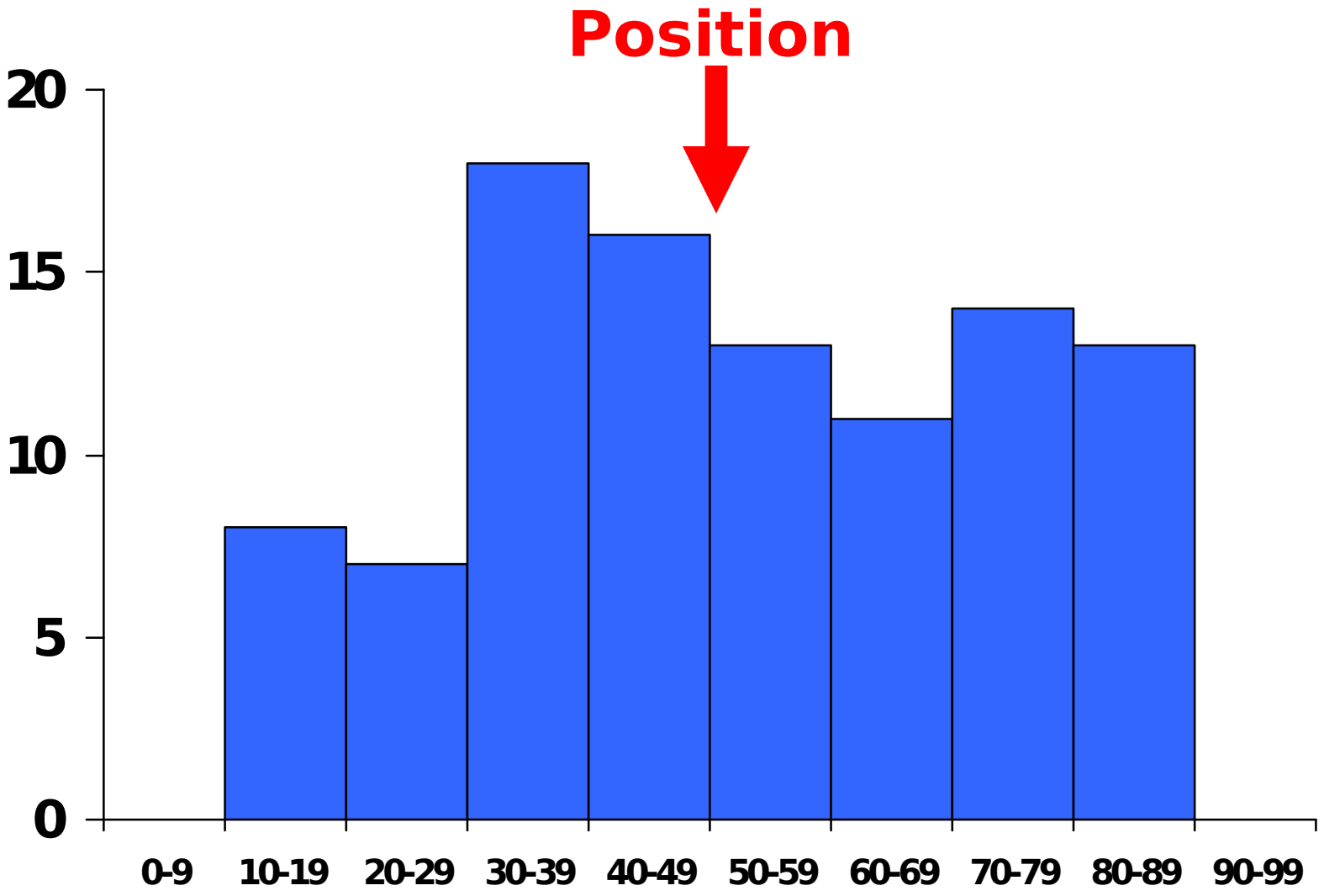
Measures of Central Tendency (MCT)

- On the scale of values of a variable there is a certain stage at which the largest number of items tend to cluster.
- Since this stage is usually in the centre of distribution, the tendency of the statistical data to get concentrated at a certain value is called **"central tendency"**
- The various methods of determining the point about which the observations tend to concentrate are called **MCT**.

- The objective of calculating MCT is to determine a single figure which may be used to represent the whole data set.

- In that sense it is an even more compact description of the statistical data than the frequency distribution.

- Since a MCT represents the entire data, it facilitates comparison within one group or between groups of data.


Position
20
15
10
5
0
0-9
10-19
20-29
30-39
40-49
50-59
60-69
70-79
80-89
90-99

Characteristics of a good MCT

A MCT is good or satisfactory if it possesses the following characteristics.

1. It should be based on all the observations
2. It should not be affected by the extreme values
3. It should be as close to the maximum number of values as possible
4. It should have a definite value
5. It should not be subjected to complicated and tedious calculations
6. It should be capable of further algebraic treatment
7. It should be stable with regard to sampling

- The most common measures of central tendency include:
  - **Arithmetic Mean**
  - **Median**
  - **Mode**
  - **Others**

1. Arithmetic Mean

A. Ungrouped Data

- The arithmetic mean is the "average" of the data set and by far the most widely used measure of central location and it is usually denoted by $\bar{x}$
- Is the sum of all the observations divided by the total number of observations.

$$\text{Mean} = \frac{\text{sum of values}}{\text{sample size}} = \frac{\Sigma\ (\text{values})}{n}$$

The Summation Notation

$\sum$ The Greek symbol sigma says "add up some items"

$\sum_{\text{STARTING HERE}}$ Below the sigma symbol is the starting point

$\sum^{\text{END}}$ Up top is the ending point

For example,

Instead of writing $x_1 + x_2 + x_3 + x_4 + x_5$

We write $\sum_{i=1}^{5} x_i$

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

$$= \frac{(x_1 + x_2 + x_3 + \ldots + x_n)}{n}$$

The heart rates for n=10 patients were as follows (beats per minute):

167, 120, 150, 125, 150, 140, 40, 136, 120, 150

What is the arithmetic mean for the heart rate of these p

$$\bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i$$

$$= \frac{1298}{10}$$

$$= 129.8 \text{ beats per minute}$$

b) Grouped data

In calculating the mean from grouped data, we assume that all values falling in a particular class interval are located at the midpoint of the interval. It is calculated as follows:

$$\bar{x} = \frac{\sum_{i=1}^{k} m_i f_i}{\sum_{i=1}^{k} f_i}$$

where,

k = the number of class intervals

m_i = the midpoint of the ith class interval

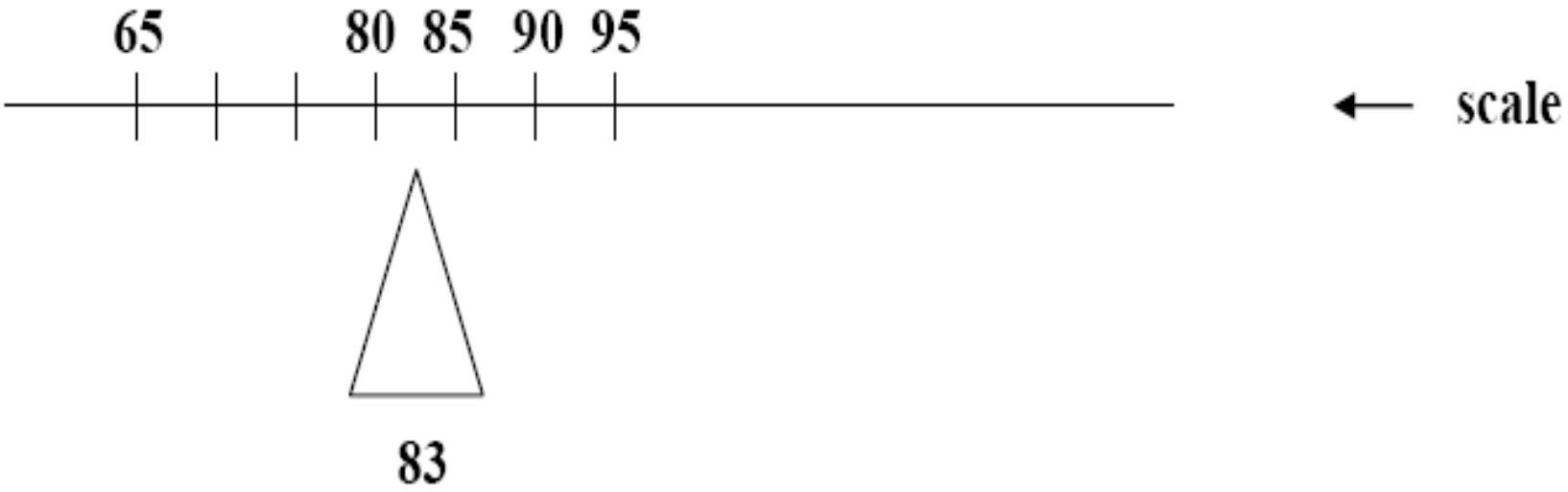
f_i = the frequency of the ith class interval

Example. Compute the mean age of 169 subjects from the grouped data.

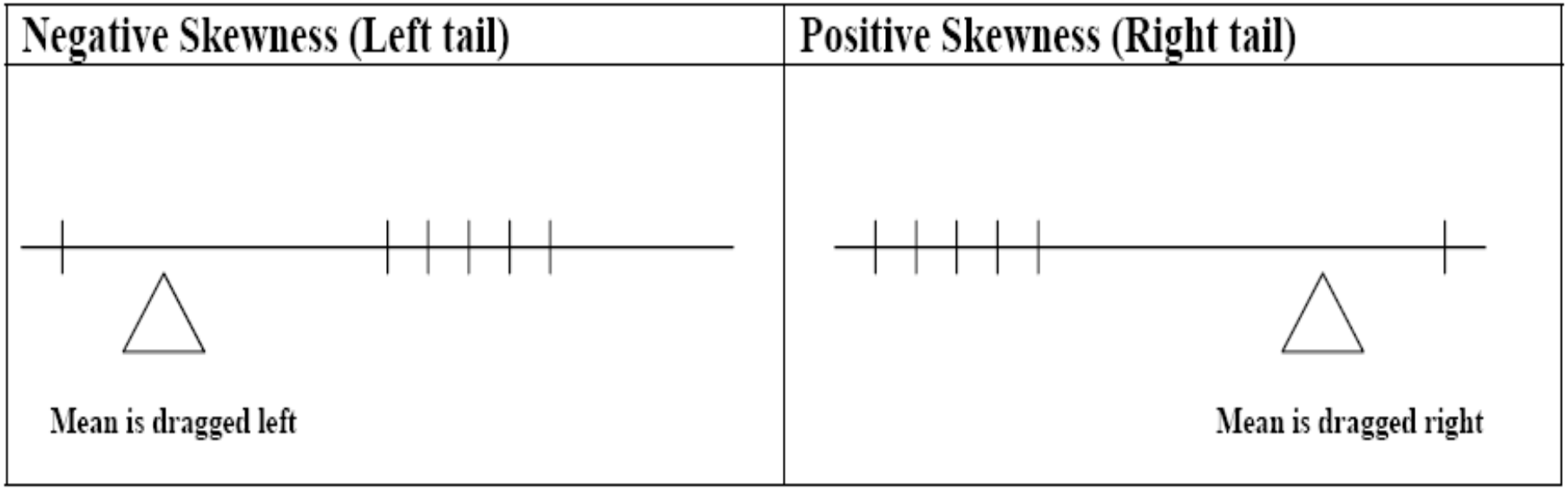
Mean = 5810.5/169 = 34.48 years

Class interval	Mid-point (m_i)	Frequency (f_i)	$m_i f_i$
10-19	14.5	4	58.0
20-29	24.5	66	1617.0
30-39	34.5	47	1621.5
40-49	44.5	36	1602.0
50-59	54.5	12	654.0
60-69	64.5	4	258.0
Total		169	5810.5

The mean can be thought of as a "balancing point", "center of gravity"



When the data are skewed, the mean is “dragged” in the direction of the skewness



- ***It is possible in extreme cases for all but one of the sample points to be on one side of the arithmetic mean & in this case, the mean is a poor measure of central location or does not reflect the center of the sample.***

Properties of the Arithmetic Mean.

- For a given set of data there is one and only one arithmetic mean (uniqueness).
- Easy to calculate and understand (simple).
- Influenced by each and every value in a data set
- Greatly affected by the extreme values.
- In case of grouped data if any class interval is open, arithmetic mean can not be calculated.

2. Median

a) Ungrouped data

- The median is the value which divides the data set into two equal parts.
- If the number of values is odd, the median will be the middle value when all values are arranged in order of magnitude.
- When the number of observations is even, there is no single middle value but two middle observations.
- In this case the median is the mean of these two middle observations, when all observations have been arranged in the order of their magnitude.

If the sample size n is ODD	$\text{median} = \frac{n+1}{2}\text{th largest value}$
If the sample size n is EVEN	$\text{median} = \text{average of}\left(\left[\frac{n}{2}\right]th, \left[\frac{n+2}{2}\right]th\right)\text{values}$

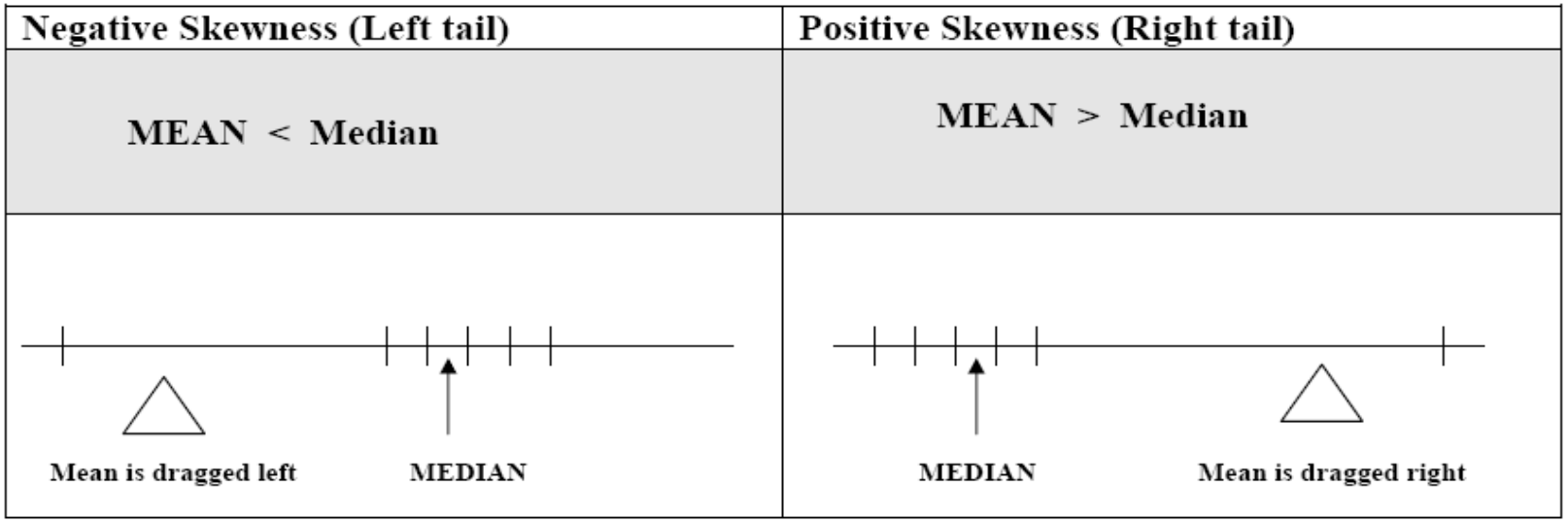
Example

- **Data, from smallest to largest, are: 1, 1, 2, 3, 7, 8, 11, 12, 14, 19, 20**
- **The sample size, n=11**
- **Median is the** $\frac{n+1}{2}$ **th largest** $= \frac{12}{2} =$ **6th largest value**
- **Thus, median value is = 8**
- **Five values are smaller than 8; five values are larger.**

Example

- Data, from smallest to largest, are: 2, 5, 5, 6, 7, 10, 15, 21, 22, 23, 23, 25
- The sample size, n=12
- Median = average of $\frac{n}{2}$th largest, $\frac{n+2}{2}$ = average of 6th and 7th largest values
- Thus, median value is = average (10, 15) = 12.5

- **The median is a better description (than the mean) of the majority when the distribution is skewed**
- **Example**
  - **Data: 14, 89, 93, 95, 96**
  - **Skewness is reflected in the outlying low value of 14**



b) Grouped data

- In calculating the median from grouped data, we assume that the values within a class-interval are evenly distributed through the interval.
- The first step is to locate the class interval in which the median is located, using the following procedure.
- Find $n/2$ and see a class interval with a minimum cumulative frequency which contains $n/2$.
- Then, use the following formula.

$$\tilde{x} = L_m + \left(\frac{\frac{n}{2} - F_c}{f_m} \right) W$$

where,

L_m = lower true class boundary of the interval containing the median

F_c = cumulative frequency of the interval just above the median class interval

f_m = frequency of the interval containing the median

W= class interval width

n = total number of observations

Example. Compute the median age of 169 subjects from the grouped data.

$n/_2 = 169/2 = 84.5$

Class interval	Mid-point (m_i)	Frequency (f_i)	Cum. freq
10-19	14.5	4	4
20-29	24.5	66	70
30-39	34.5	47	117
40-49	44.5	36	153
50-59	54.5	12	165
60-69	64.5	4	169
Total		169	

- $n/_2$ = 84.5 = in the 3rd class interval
- Lower limit = 29.5, Upper limit = 39.5
- Frequency of the class = 47
- $(n/_2 - f_c)$ = 84.5-70 = 14.5

- **Median = 29.5 + (14.5/47)10 = 32.58 ≈ 33**

Properties of the median

- There is only one median for a given set of data (uniqueness)
- The median is easy to calculate
- Median is a positional average and hence it is insensitive to very large or very small values
- Median can be calculated even in the case of open end intervals
- It is determined mainly by the middle points and less sensitive to the remaining data points (weakness).

Quartiles

- Just as the median is the value above and below which lie half the set of data, one can define measures (above or below) which lie other fractional parts of the data.
- The median divides the data into two equal parts
- If the data are divided into four equal parts, we speak of quartiles.

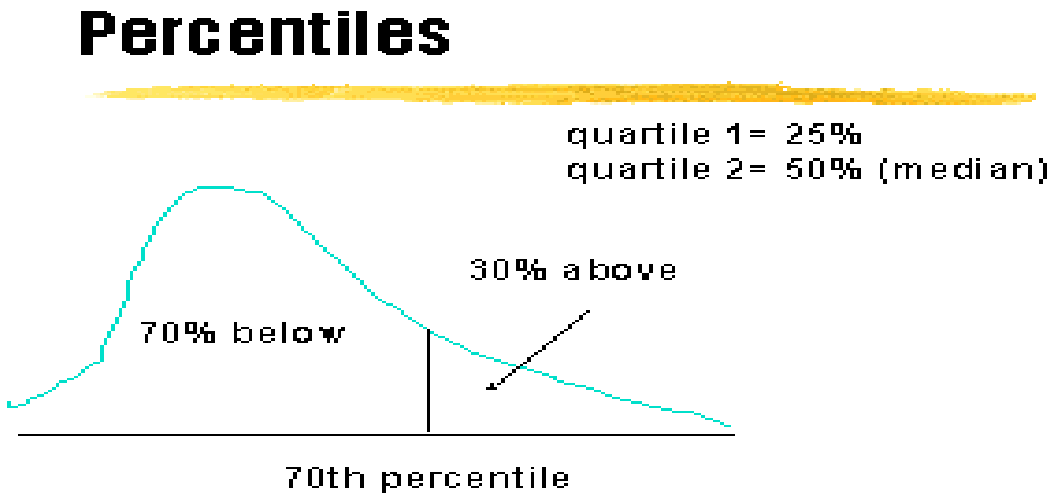
a) The first quartile (Q_1): 25% of all the ranked observations are less than Q_1.

b) The second quartile (Q_2): 50% of all the ranked observations are less than Q_2. The second quartile is the median.

c) The third quartile (Q_3): 75% of all the ranked observations are less than Q_3.

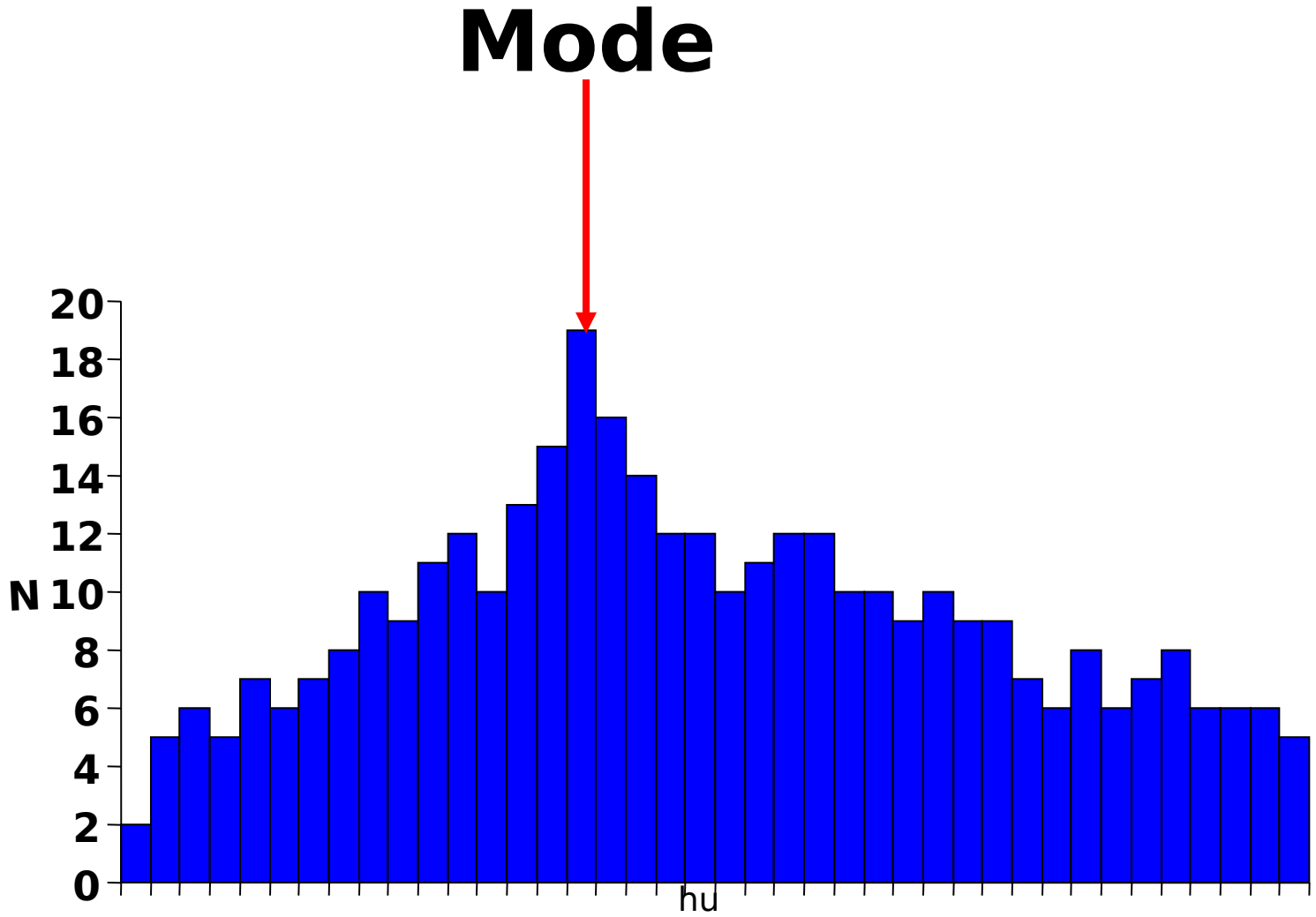
Percentiles

- Simply divide the data into 100 pieces.
- Percentiles are less sensitive to outliers and not greatly affected by th



3. Mode

- The mode is the most frequently occurring value among all the observations in a set of data.
- It is not influenced by extreme values.
- It is possible to have more than one mode or no mode.
- It is not a good summary of the majority of the data.


Mode
20
18
16
14
12
N 10
8
6
4
2
0
hu

a) Ungrouped data

- It is a value which occurs most frequently in a set of values.
- If all the values are different there is no mode, on the other hand, a set of values may have more than one mode.

- **Example**
  - **Data are: 1, 2, 3, 4, 4, 4, 4, 5, 5, 6**
  - **Mode is 4 “Unimodal”**
- **Example**
  - **Data are: 1, 2, 2, 2, 3, 4, 5, 5, 5, 6, 6, 8**
  - **There are two modes – 2 & 5**
  - **This distribution is said to be “bi-modal”**
- **Example**
  - **Data are: 2.62, 2.75, 2.76, 2.86, 3.05, 3.12**
  - **No mode, since all the values are different**

b) Grouped data

- To find the mode of grouped data, we usually refer to the modal class, where the modal class is the class interval with the highest frequency.
- If a single value for the mode of grouped data must be specified, it is taken as the mid-point of the modal class interval.

$$\hat{x} = L_m + w\left(\frac{f_2}{f_0 + f_2}\right)$$

where

L - Lower boundary of the Modal class

f_0 – The frequency of the class next below the modal class in value

f_2 – the frequency of the class next above the modal class in value

w – length of the interval of the modal class

Example

Interval of Values ("Class")	Frequency, f
31-40	1
41-50	2
51-60	5
61-70	15
71-80	25
81-90	20
91-100	12

- The modal class is the interval of values 71-80

Properties of mode

- It is not affected by extreme values
- It can be calculated for distributions with open end classes
- Often its value is not unique
- The main drawback of mode is that often it does not exist

4. **Geometric mean (GM)**

- Mainly used in many types of laboratory data, specifically data in the form of concentrations of one substance in another
- Example: the minimum inhibitory concentration of penicillin in urine for *N. gonorrhoeae* in 71 patients

(µg/ml)	Frequency	(µg/ml)	Frequency
0.03125	**21**	**0.250**	**19**
0.0625	**6**	**0.50**	**17**
0.1250	**8**	**1.0**	**3**

If $x_1, x_2, \ldots, x_n$ are n positive observed values, then

$$GM = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

and

$$\log GM = \frac{\sum_{i=1}^{n} \log x_i}{n}.$$

The geometric mean is generally used with data measured on a logarithmic scale, such as titers of anti-neutrophil immunoglobulin G.

Example:

$\log GM = [21\log(0.03125) + 6\log(0.0625) + 8\log(0.125) + 19\log(0.25) + 17\log(0.5) + 3\log(1.0)]/74 = -0.846$

The GM = the antilogarithm of -0.846 = 0.143

5. **Harmonic mean (HM)**

- Just as the geometric mean is based on an arithmetic mean of logarithms, so is the harmonic mean based on arithmetic mean of the reciprocals.
- Pertains to rates and time
- We define it as the reciprocal of the arithmetic mean of the reciprocal of the given numbers.

If the given numbers are $x_1, x_2, \ldots, x_n$, then

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

6. **Weighted mean (WM)**

- In a weighted mean, separate outcomes have separate influences.
- The influence attached to an outcome is the weight.
- Familiar is the calculation of a course grade as a weighted average of scores on separate outcomes.

Example:

Outcome, X_i is	**Weight (Influence) w_i is**
Homework #1 = $X_1 = 96$	$w_1 = 20$ **(20% of grade)**
Homework #2= $X_2 = 84$	$w_2 = 20$
Homework #3 = $X_3 = 65$	$w_3 = 20$
Midterm Exam = $X_4 = 73$	$w_4 = 10$
Final Exam = $X_5 = 94$	$w_5 = 30$

$$\text{weighted mean} = \frac{\sum(\text{weight})(\text{score})}{\sum(\text{weight})}$$

$$= \frac{\sum_{i=1}^{n}(w_i)(X_i)}{\sum(w_i)}$$

$$= \frac{(20)(96)+(20)(84)+(20)(65)+(10)(73)+(30)(94)}{(20)+(20)+(20)+(10)+(30)}$$

$$= 84.5$$

hu

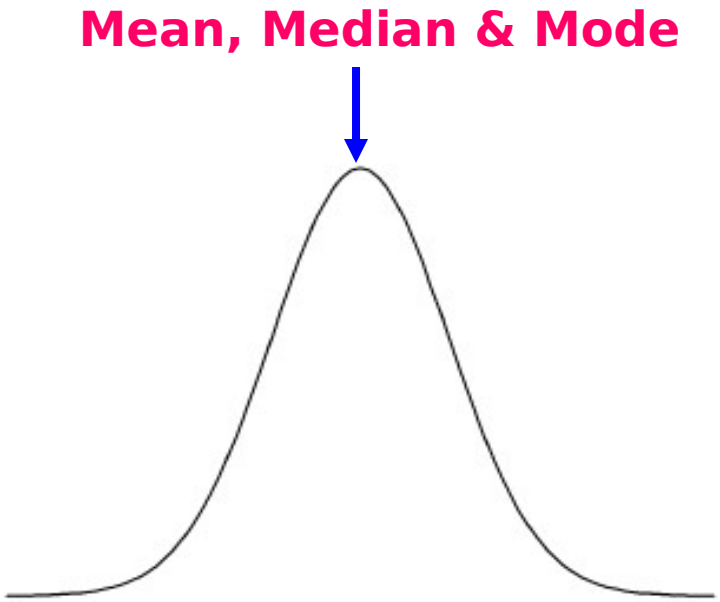
Which measure of central tendency is best with a given set of data?

- Two factors are important in making this decisions:
  - The scale of measurement (type of data)
  - The shape of the distribution of the observations

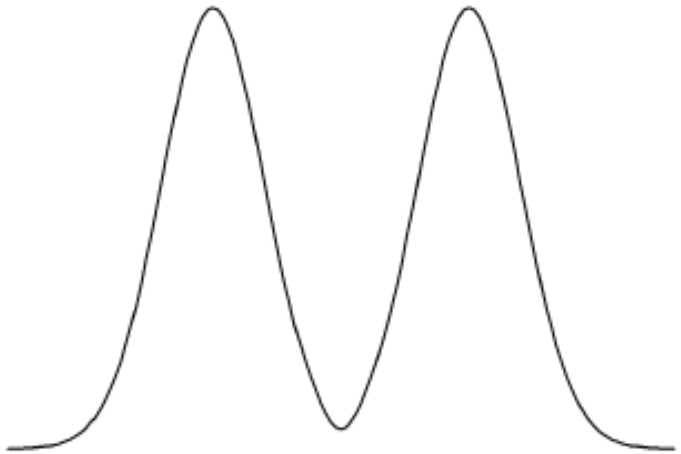
- The mean can be used for discrete and continuous data
- The median is appropriate for discrete and continuous data as well, but can also be used for ordinal data
- The mode can be used for all types of data, but may be especially useful for nominal and ordinal measurements
- For discrete or continuous data, the “modal class” can be used

- The geometric mean is used primarily for observations measured on a logarithmic scale.
- Harmonic mean is a suitable MCT when the data pertains to rates and time.
- Weighted mean is commonly used in the calculation of mean for different outcomes.

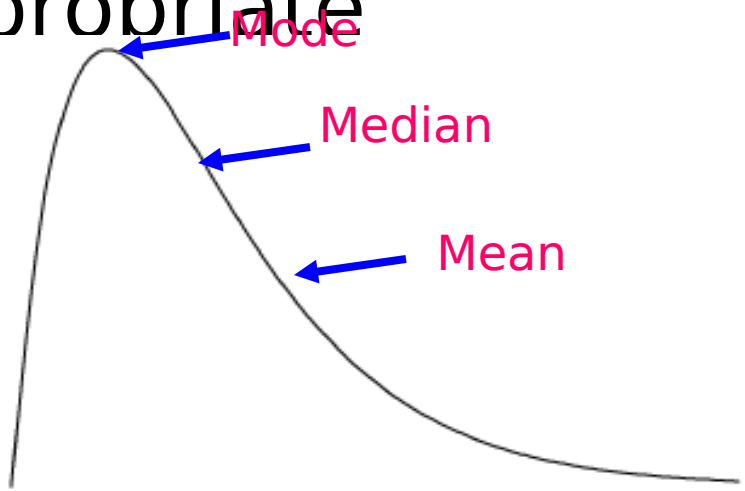
(a) Symmetric and unimodal distribution — Mean, median, and mode should all be approximately the same



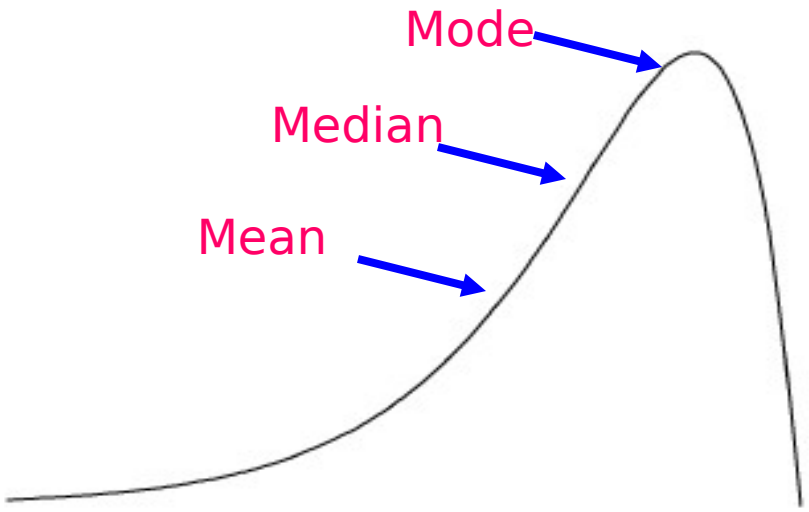
(b) Bimodal — Mean and median should be about the same, but may take a value that is unlikely to occur; two modes might be best



(c) Skewed to the right (positively skewed) —Mean is sensitive to extreme values, so median might be more appropriate



(d) Skewed to the left (negatively skewed) — Same as (c)



Measures of Dispersion

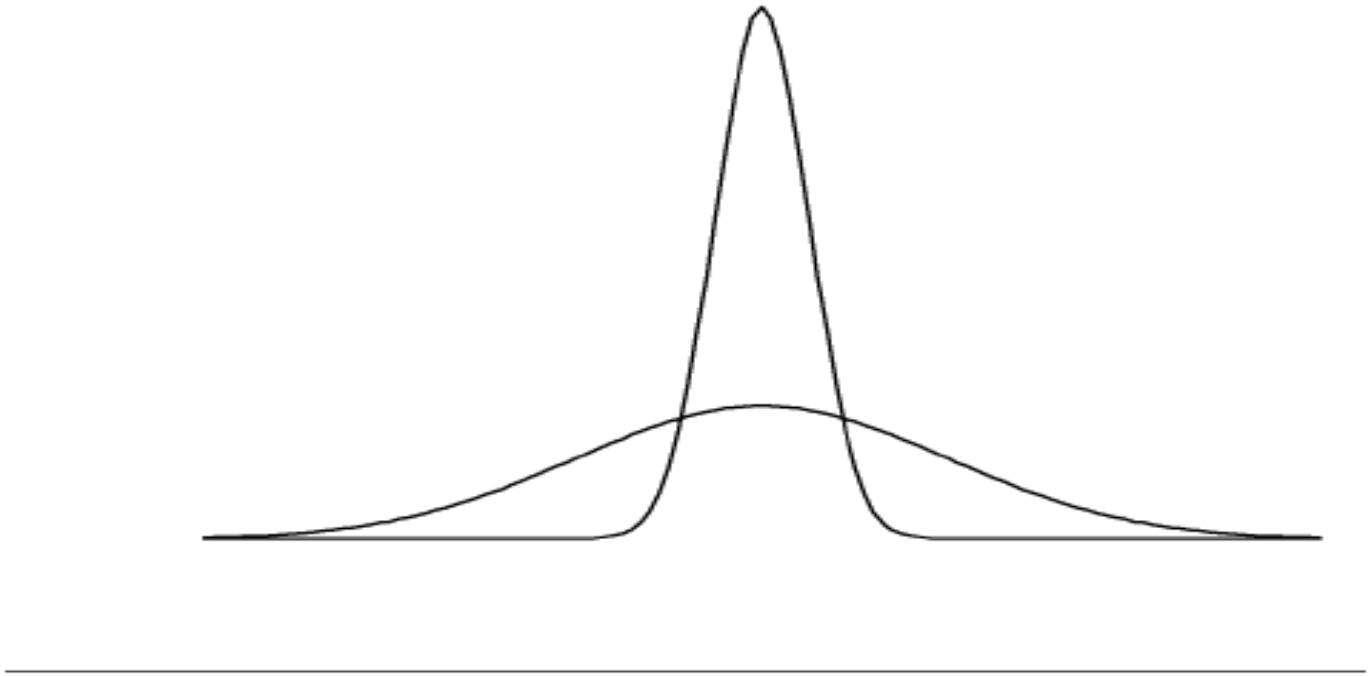
Consider the following two sets of data:

A: 177 193 195 209 226
Mean = 200

B: 192 197 200 202 209
Mean = 200

Two or more sets may have the same mean and/or median but they may be quite different.

These two distributions have the same mean,



- MCT are not enough to give a clear understanding about the distribution of the data.

- We need to know something about the variability or spread of the values — whether they tend to be clustered close together, or spread out over a broad range

Measures of Dispersion

- **Measures that quantify the variation or dispersion of a set of data from its central location**
- **Dispersion refers to the variety exhibited by the values of the data.**
- **The amount may be small when the values are close together.**
- ***If all the values are the same, no dispersion***

Measures of Dispersion

Other synonymous term:

- “Measure of Variation”
- “Measure of Spread”
- “Measures of Scatter”

- **Measures of dispersion include:**
  - **Range**
  - **Inter-quartile range**
  - **Variance**
  - **Standard deviation**
  - **Coefficient of variation**
  - **Standard error**
  - **Others**

1. Range (R)

- The difference between the largest and smallest observations in a sample.

- Range = Maximum value – Minimum value

- **Example –**
  - **Data values: 5, 9, 12, 16, 23, 34, 37, 42**
  - **Range = 42-5 = 37**
- Data set with higher range exhibit more variability

Properties of range

- It is the simplest crude measure and can be easily understood
- It takes into account only two values which causes it to be a poor measure of dispersion
- Very sensitive to extreme observations
- The larger the sample size, the larger the range

2. Interquartile range (IQR)

- Indicates the spread of the middle 50% of the observations, and used with median

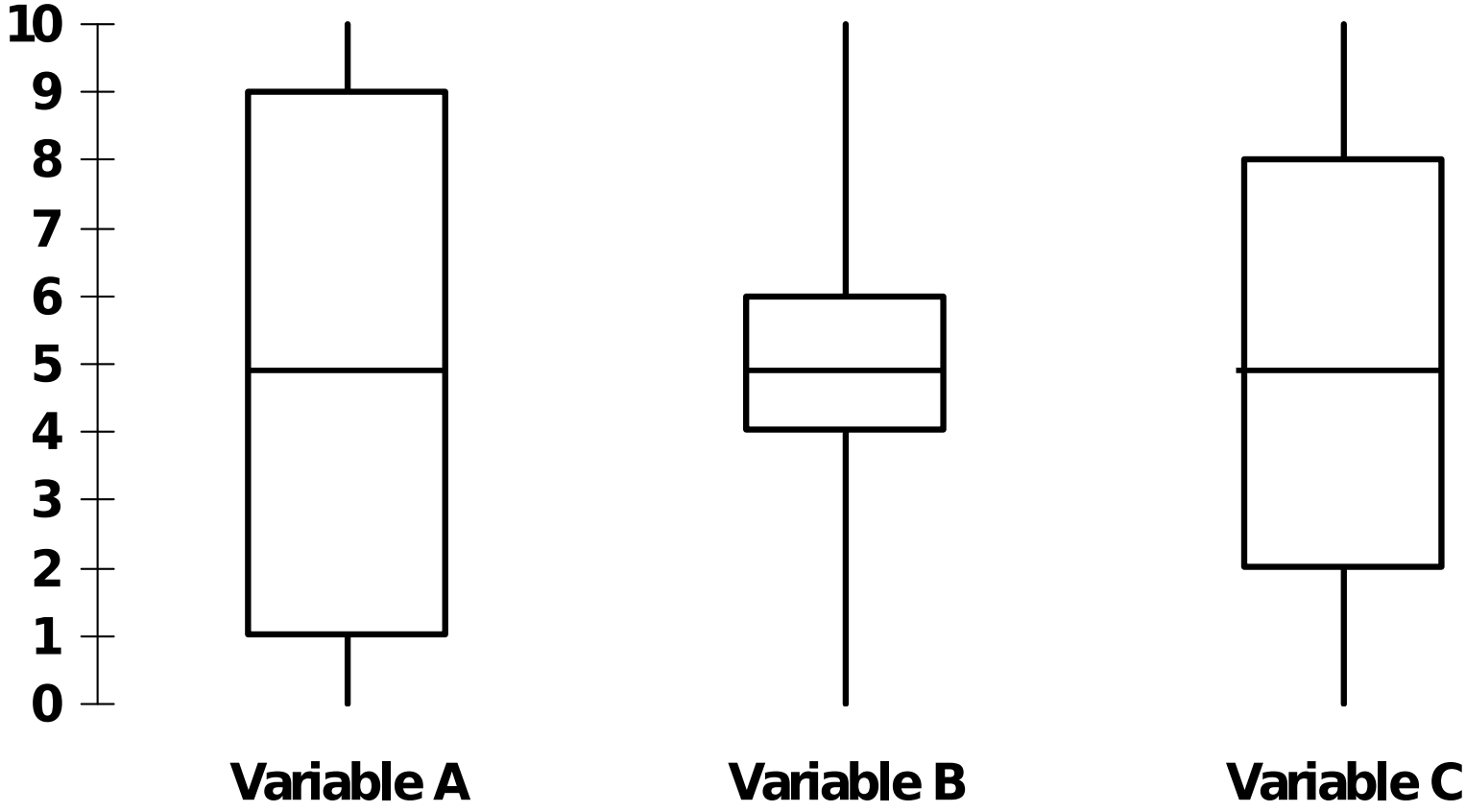
$$IQR = Q_3 - Q_1$$

- **Example:** Suppose the first and third quartile for weights of girls 12 months of age are 8.8 Kg and 10.2 Kg, respectively.

IQR = 10.2 Kg – 8.8 Kg

i.e., 50% of the infant girls weigh between 8.8 and 10.2 Kg.

The two quartiles (Q3 &Q1) form the basis of the Box-and-Whiskers Plots — Variables A, B, C



Properties of IQR:

- It is a simple and versatile measure
- It encloses the central 50% of the observations
- It is not based on all observations but only on two specific values
- It is important in selecting cut-off points in the formulation of clinical standards
- Since it excludes the lowest and highest 25% values, it is not affected by extreme values
- Less sensitive to the size of the sample

3. Quartile deviation (QD)

$$QD = \frac{Q_3 - Q_1}{2}$$

4. **Coefficient of quartile deviation (CQD)**

- CQD = $\dfrac{Q_3 - Q_1}{Q_3 + Q_1}$

- CQD is an absolute quantity (unitless) and is useful to compare the variability among the middle 50% observations.

5. Mean deviation (MD)

- Mean deviation is the average of the absolute deviations taken from a central value, generally the mean or median.
- Consider a set of n observations x_1, x_2, ..., x_n. Then:

$$MD = \frac{1}{n}\sum_{i=1}^{n}|x_i - A|$$

- 'A' is a central value (arithmetic mean or median).

Properties of mean deviation:

- MD removes one main objection of the earlier measures, that *it involves each value*
- It is not affected much by extreme values
- Its main drawback is that algebraic negative signs of the deviations are ignored which is mathematically unsound

6. Variance (σ^2, s^2)

- The main objection of mean deviation, that the negative signs are ignored, is removed by taking the square of the deviations from the mean.

- The variance is the average of the squares of the deviations taken from the mean.

- It is squared because the sum of the deviations of the individual observations of a sample about the sample mean is always 0

$$0 = \sum (x_i - \bar{x})$$

- The variance can be thought of as an average of squared deviations

- ***Variance*** is used to measure the dispersion of values relative to the mean.
- When values are close to their mean (narrow range) the dispersion is less than when there is scattering over a wide range.
  - **Population variance = σ^2**
  - **Sample variance = S^2**

a) Ungrouped data

- Let $X_1, X_2, \ldots, X_N$ be the measurement on N population units, then:

$$\sigma^2 = \frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N} \quad \text{where}$$

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} \quad \text{is the population mean.}$$

A sample variance is calculated for a sample of individual values (X_1, X_2, … X_n) and uses the sample mean (e.g. $\overline{X}$) rather than the population mean μ.

$$S^2 = \frac{\text{sum of (value} - \text{sample mean)}^2}{\text{sample size} - 1}$$

$$= \frac{\sum (\text{value} - \text{sample mean})^2}{n - 1}$$

$$= \frac{\sum_{i=1}^{n} (X - \overline{X})^2}{n - 1}$$

Example: Areas of sprayable surfaces with DDT from a sample of 15 houses are as follows (m2) : 101,105,110,114,115,124,125, 125, 130,133,135,136,137,140,145

- Find the variance and standard deviation of the above distribution. The mean of the sample is 125 m2.
- Variance (sample) = $s2 = \Sigma(xi - x)2/n-1$
- = {(101-125)2 +(105-125)2 + ….(145-125)2 } / (15-1)
- = 2502/14 = 178.71 (square metres)2,Hence, the standard deviation = 178.71 = 13.37 m2.

Degrees of freedom

- Which is estimate or make difference about population parameter based on sample data.
- In computing the variance there are (n-1) degrees of freedom because only (n-1) of the deviations are independent from each other
- The last one can always be calculated from the others automatically.
- This is because the sum of the deviations from their mean (Xi-Mean) must add to zero.

b) Grouped data

$$S^2 = \frac{\sum_{i=1}^{k} (m_i - \bar{x})^2 f_i}{\sum_{i=1}^{k} f_i - 1}$$

where

m_i = the mid-point of the ith class interval

f_i = the frequency of the ith class interval

$\bar{x}$ = the sample mean

k = the number of class intervals

Properties of Variance:

- The main disadvantage of variance is that its unit is the square of the unite of the original measurement values
- The variance gives more weight to the extreme values as compared to those which are near to mean value, because the difference is squared in variance.
- The drawbacks of variance are overcome by the standard deviation.

7. Standard deviation (σ, s)

- **It is the square root of the variance.**
- **This produces a measure having the same scale as that of the individual values.**

$$\sigma = \sqrt{\sigma^2} \text{ and } S = \sqrt{S^2}$$

- Following are the survival times of n=11 patients after heart transplant surgery.

- The survival time for the “ith” patient is represented as Xi for i= 1, ..., 11.

- Calculate the sample variance and SD.

Patient Identifier, "i"	Survival (days), X_I	Mean for sample, $\overline{X}$	Deviation , $(X_i - \overline{X})$	Squared deviation $(X_i - \overline{X})^2$
1	135	161	-26	676
2	43	161	-118	13924
3	379	161	218	47524
4	32	161	-129	16641
5	47	161	-114	12996
6	228	161	67	4489
7	562	161	401	160801
8	49	161	-112	12544
9	59	161	-102	10404
10	147	161	-14	196
11	90	161	-71	5041
TOTAL	1771		0	285236

- $s^2 = \dfrac{\sum_{i=1}^{11} (X_i - \overline{X})^2}{n-1} = \dfrac{285236}{10} = 28523.6 \ \textbf{days}^2$
- $s = \sqrt{s^2} = \sqrt{28523.6} = 168.89 \ \textbf{days}$

Example. Compute the variance and SD of the age of 169 subjects from the grouped data.

Mean = 5810.5/169 = 34.48 years

S^2 = 20199.22/169-1 = 120.23

SD = √S2 = √120.23 = 10.96

Class interval	(m_i)	(f_i)	(m_i-Mean)	(m_i-Mean)^2	(m_i-Mean)^2 fi
10-19	14.5	4	-19.98	399.20	1596.80
20-29	24.5	66	-9-98	99.60	6573.60
30-39	34.5	47	0.02	0.0004	0.0188
40-49	44.5	36	10.02	100.40	3614.40
50-59	54.5	12	20.02	400.80	4809.60
60-69	64.5	4	30.02	901.20	3604.80
Total		169		1901.20	20199.22

Properties of SD

- The SD has the advantage of being expressed in the same units of measurement as the mean

- SD is considered to be the best measure of dispersion and is used widely because of the properties of the theoretical normal curve.

- However, if the units of measurements of variables of two data sets is not the same, then there variability can’t be compared by comparing the values of SD.

SD Vs Standard Error (SE)

- **SD** describes the variability among individual values in a given data set
- **SE** is used to describe the variability among separate sample means obtained from one sample to another

$$SE(\overline{X}) = \frac{SD}{\sqrt{n}}$$

- We interpret SE to mean that another similarly conducted study may give a mean that may lie between $\overline{X} \pm$ SE.

Standard Error

- **SD is about the variability of individuals**

- **SE is used to describe the variability in the means of repeated samples taken from the same population.**

- *For example, imagine 5,000 samples, each of the same size n=11. This would produce 5,000 sample means. This new collection has its own pattern of variability. We describe this new pattern of variability using the SE, not the SD.*

Example: The heart transplant surgery

n=11, SD=168.89, Mean=161 days

- What happens if we repeat the study? What will our next mean be? Will it be close? How different will it be? Focus here is on the generalizability of the study findings.
- The behavior of mean from one replication of the study to the next replication is referred to as the sampling distribution of mean.
- We can also have sampling distribution of the median or the SD

$$SE(\overline{X}) = \frac{SD}{\sqrt{n}} = \frac{168.89}{\sqrt{11}} = 50.9$$

- We interpret this to mean that a similarly conducted study might produce an average survival time that is near 161 days, ±50.9 days.

8. Coefficient of variation (CV)

- When two data sets have different units of measurements, or their means differ sufficiently in size, the CV should be used as a measure of dispersion.
- It is the best measure to compare the variability of two series of sets of observations.
- Data with less coefficient of variation is considered more consistent.

- CV is the ratio of the SD to the mean multiplied by 100.

$$CV = \frac{S}{\bar{x}} \times 100$$

	SD	Mean	CV (%)
SBP	15mm	130mm	11.5
Cholesterol	40mg/dl	200mg/dl	20.0

- "Cholesterol is more variable than systolic blood pressure"

NOTE:

- The range often appears with the median as a numerical summary measure
- The IQR is used with the median as well
- The SD is used with the mean
- For nominal and ordinal data, a table or graph is often more effective than any numerical summary measure