

Lecture #3

Descriptive Statistics

- Measures of location (central tendency)
 - Mode
 - Median
 - Mean
 - Measures of dispersion, variability
 - Range
 - IQR
 - Variance
 - SD
 - Measures of position
 - Quintiles
 - Deviation score
 - Z score
 - Measures of symmetry & peakedness
 - Skewness
 - Kurtosis
 - Other measures
 - Weighted average
 - Geometric mean
-

Variables and distributions

- A variable is a characteristic that changes or takes on different values (e.g., age, SBP).
 - A distribution consists of values of a characteristic and the frequency of their occurrence.
 - Distributions of numbers can be summarized with numbers (called statistics or parameters).
-

Parameters and Statistics

A quantity such as a measure of central tendency or dispersion for an entire population is a **parameter**

An **estimate** of a population parameter is a **statistic**

Parameters from **P**opulations

Statistics from **S**amples

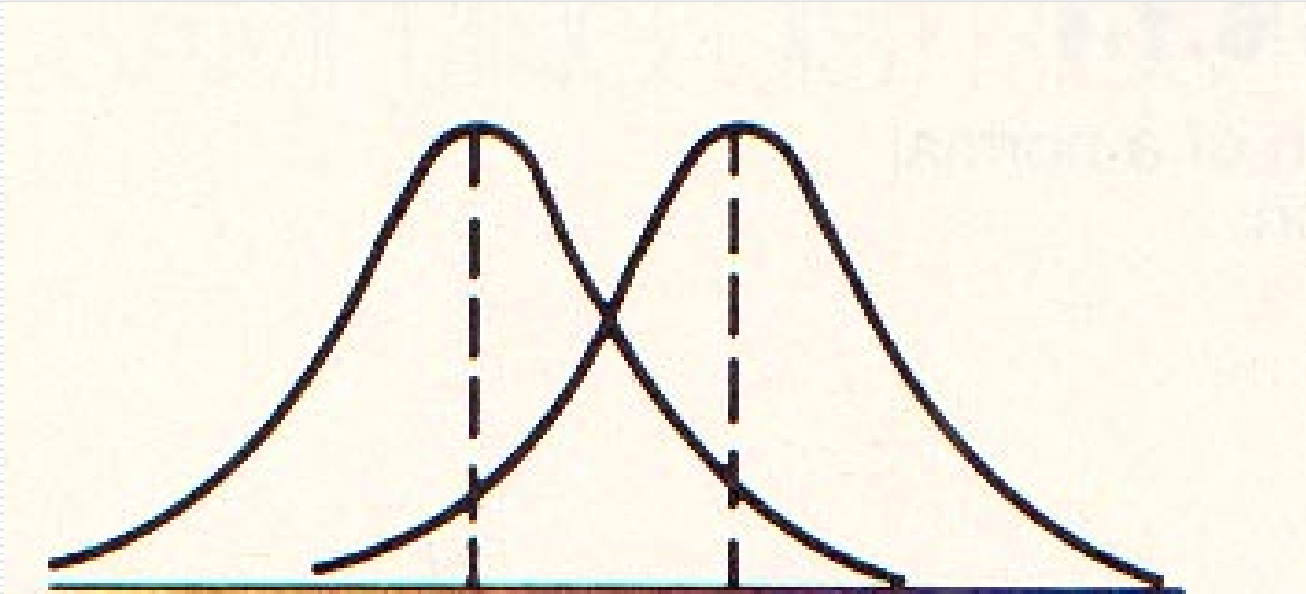
Parameters are represented by Greek letters, while statistics are represented by Latin letters.

e.g Population mean = μ (Greek mu)

Sample mean = \bar{X} , (X bar)

Central tendency

- It refers to an average score in a set of scores; or the middle of a distribution



Central tendency

- Trying to specify a “representative” value for a set of score
 - Typical measures are
 - mode
 - median
 - mean
-

Mode

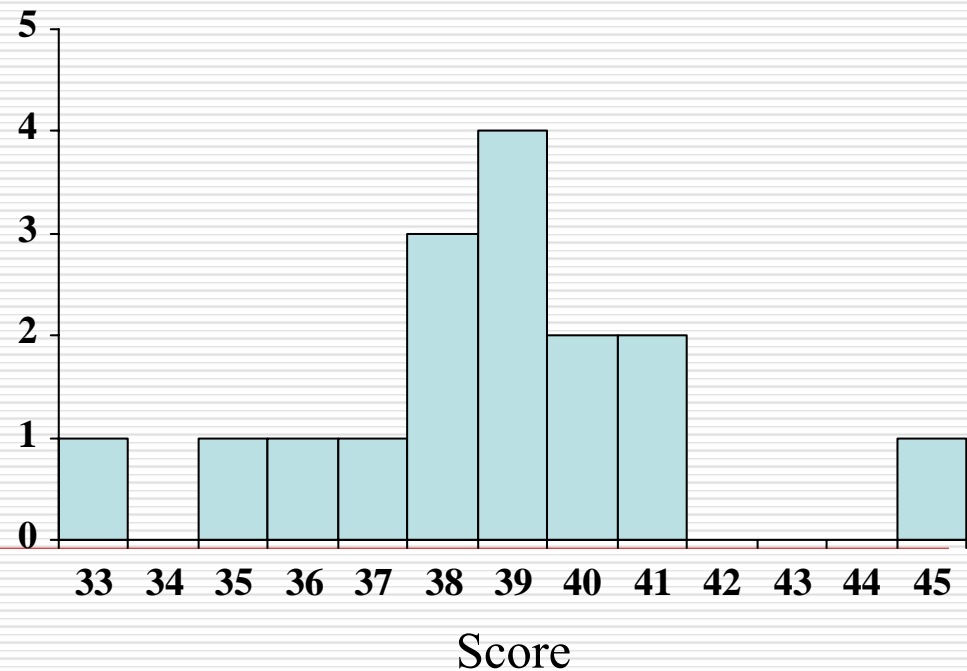
- the most frequently occurring score value
- The most easily calculated index of central tendency
- corresponds to the highest point on a frequency distribution, graph

For a given sample

$N=16$:

33 35 36 37 38 38 38
39 39 39 39 40 40 41
41 45

The mode = 39



Mode

- The mode is not sensitive to extreme scores.

For a given sample

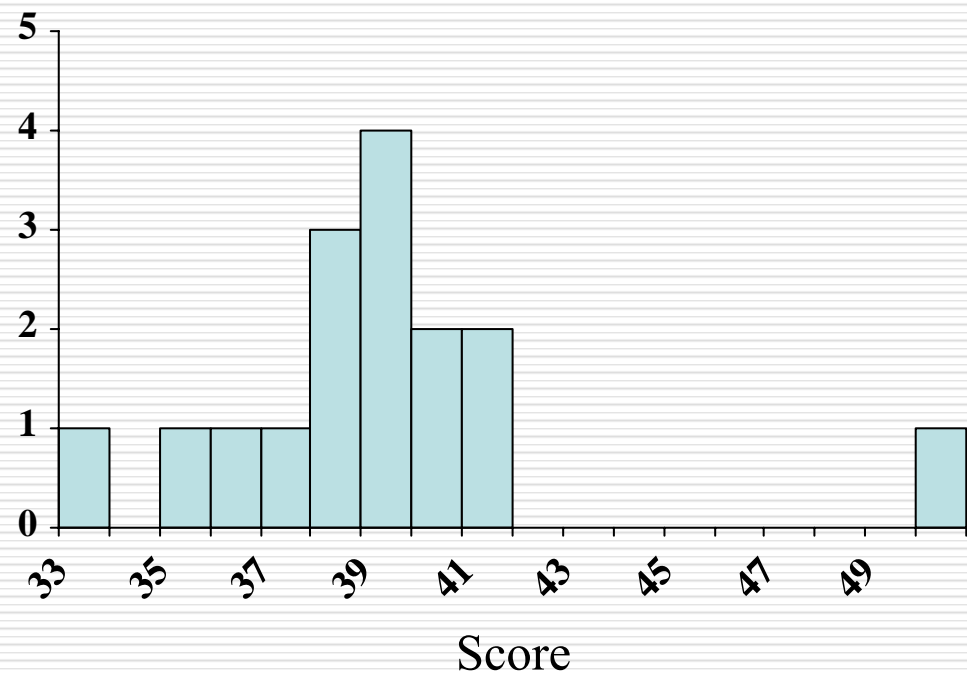
$N=16$:

33 35 36 37 38 38 38

39 39 39 39 40 40 41

41 50

The mode = 39



Mode

- a distribution may have more than one mode, bimodal distribution

For a given sample

N=16:

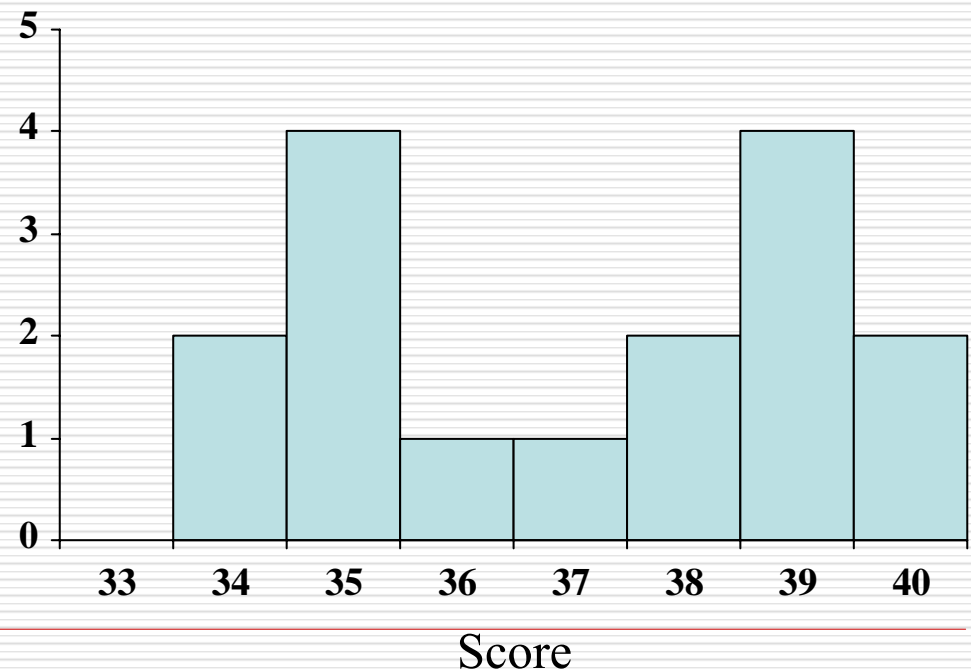
34 34 35 35 35 35 36

37 38 38 39 39 39 39

40 40

The modes = 35 and

39



Mode

- there may be no unique mode, as in the case of a rectangular distribution

For a given sample

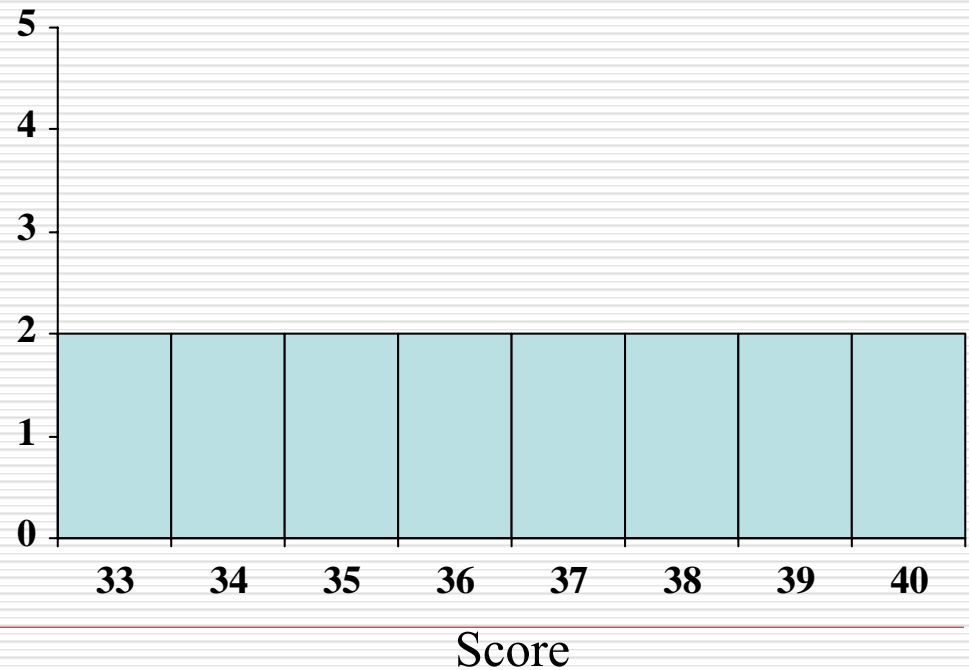
$N=16$:

33 33 34 34 35 35 36

36 37 37 38 38 39 39

40 40

No unique mode



Mode

- The only measure of location that could be used for categorical data

e.g. the most frequently ordered meal by patients admitted in gynae. ward

Chicken, meat, or fish

Median

- The middle number in a distribution
 - Divide the distribution into 50% below and 50% above it
-

Median

- Sample of $n = 5$
- Rank in order of magnitude

3	5	2	4	2
---	---	---	---	---

$$X_1 = 2$$

$$X_2 = 2$$

$$X_3 = 3$$

$$X_4 = 4$$

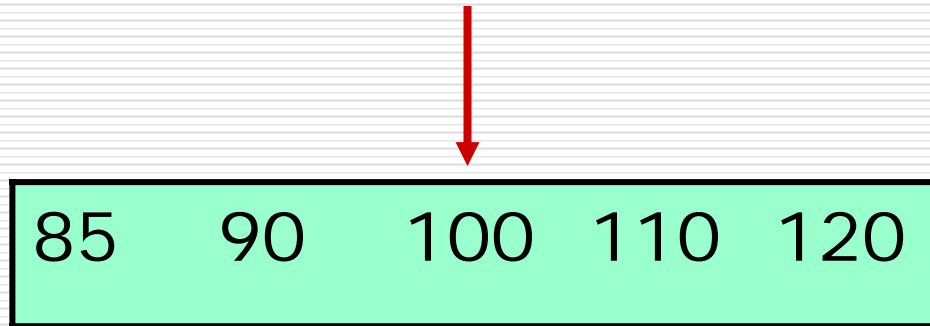
$$X_5 = 5$$

Select the middle number



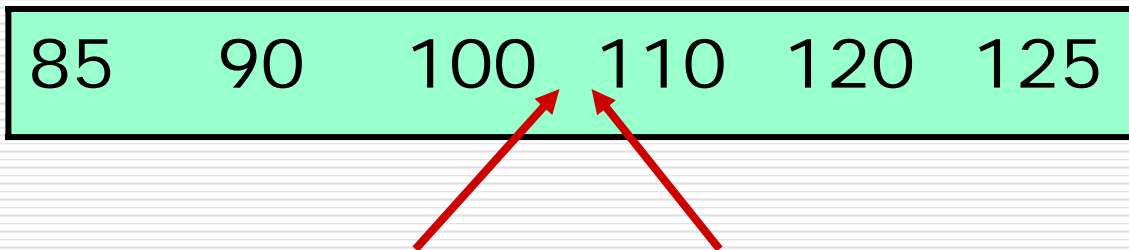
Median

- For **odd** number of observations, median = $(n+1)/2 =$ value of **kth** observation
- Sample size = 5, median is
- $(5+1)/2 =$ value of **3rd** observation



Median

- For **even** number of observations, median is the average of the middle 2 scores = midpoint of $(n)/2$ & $(n/2)+1$
- Sample size = 6, median is average of
- $(6/2)$ & $(6/2)+1 =$ value of **3rd** and **4th** observations and



- Median = $100 + 110 / 2 = 105$
-

Median

- ❑ Is not affected by extreme values or outliers;
 - ❑ The smallest absolute error
 - ❑ Could be used with data measured on ordinal scale
 - ❑ Could be estimated from a frequency table
 - ❑ Best with skewed distribution
 - ❑ But not reliable as much as the mean, no account for all observations
-

Mean

- this is what people usually have in mind when they say “average”
- the sum of the scores divided by the number of scores

For a sample:

$$\bar{X} = \frac{\sum X}{n}$$

For a population:

$$\mu = \frac{\sum X}{N}$$

Changing the value of a single score may not affect the mode or median, but it will affect the mean.

Example:

$$2 = X_1$$

$$3 = X_2$$

$$5 = X_3$$

$$2 = X_4$$

$$4 = X_5$$

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

$$= \frac{X_1 + X_2 + X_3 + X_4 + X_5}{5}$$

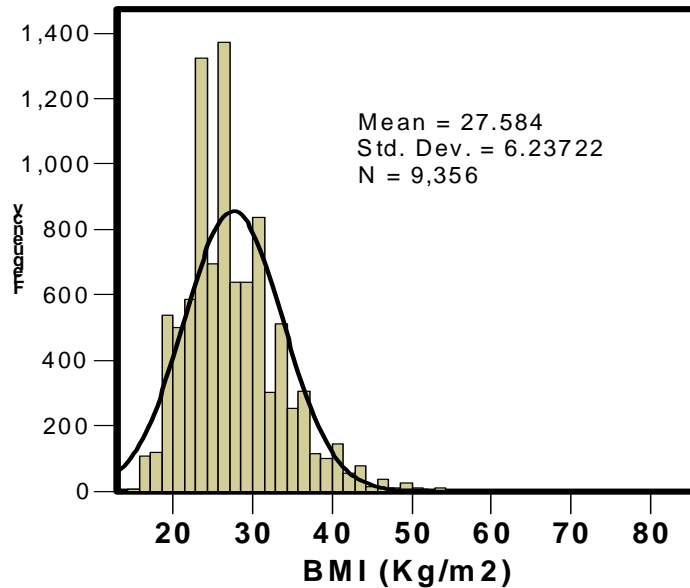
$$= \frac{2 + 3 + 5 + 2 + 4}{5}$$

$$= \frac{16}{5}$$

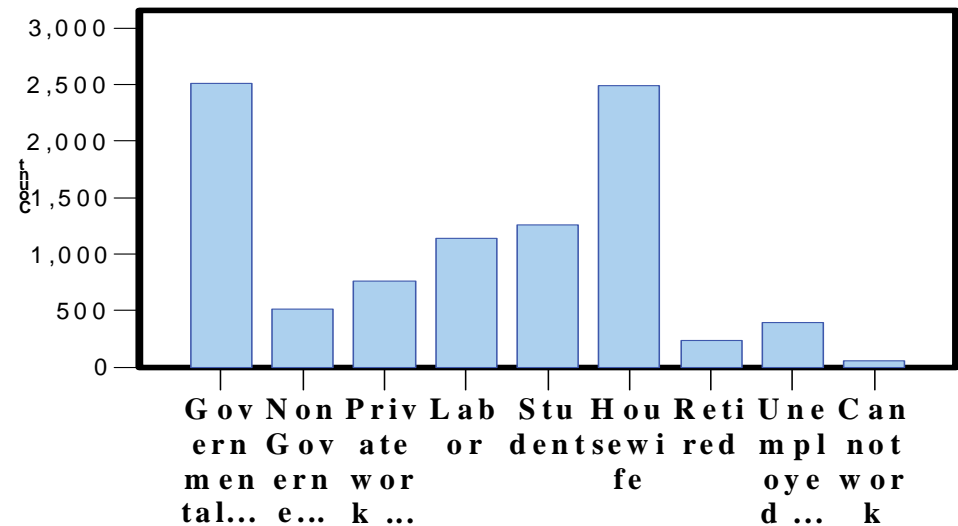
$$= 3.2$$

Mean

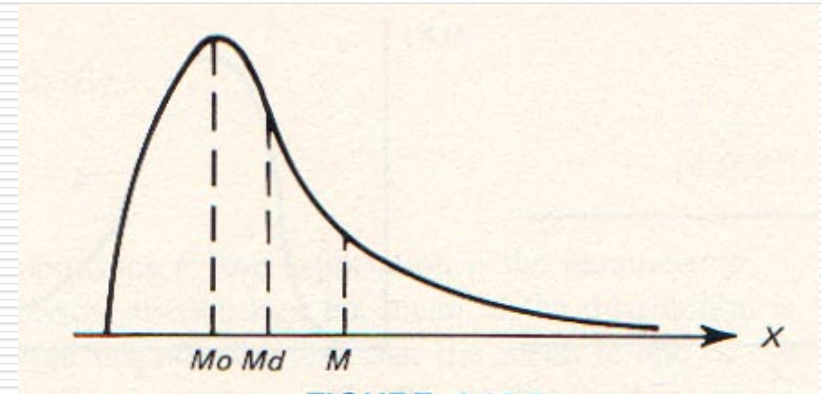
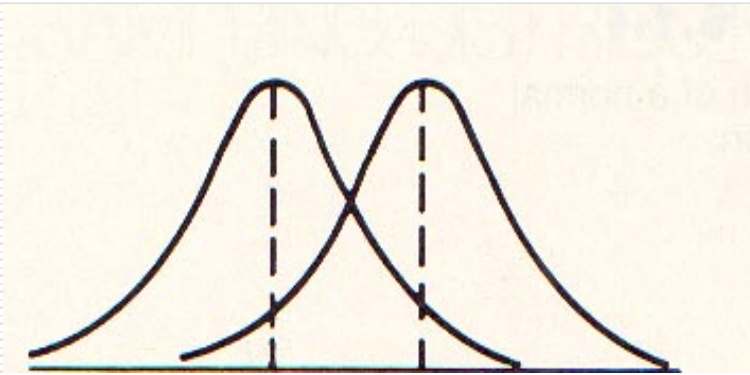
In many cases the mean is the preferred measure of central tendency, both as a description of the data and as an estimate of the parameter.



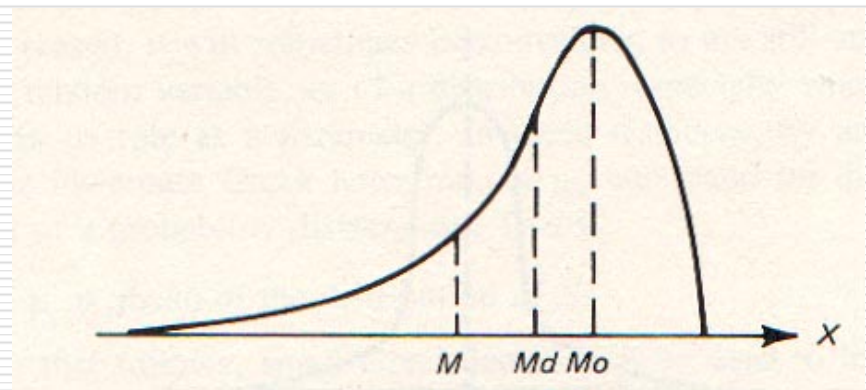
Mean to be meaningful should be used with data measured on an interval or ratio scale; numerical data only



Mean



The mean is sensitive to extreme scores and is appropriate for more symmetrical distributions.

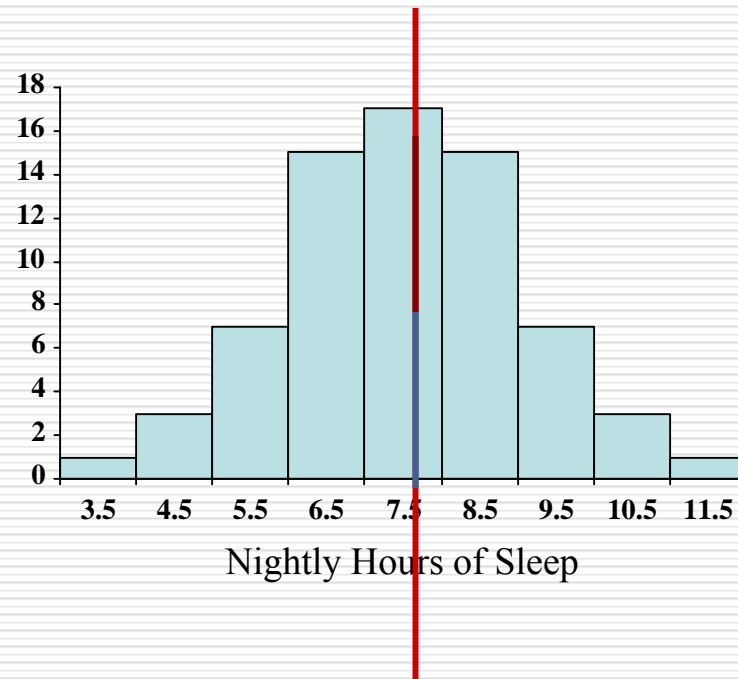


Arithmetic mean

- ❑ Most reliable and commonly used statistic & estimate of parameter
 - ❑ Consider all values in calculation
 - ❑ Unique value, easy calculation
 - ❑ Affected by extreme value, outliers
 - ❑ Not suitable for ordinal data
-

Symmetry

- a symmetrical distribution exhibits no skewness
- in a symmetrical distribution the **Mean** = **Median** = **Mode**



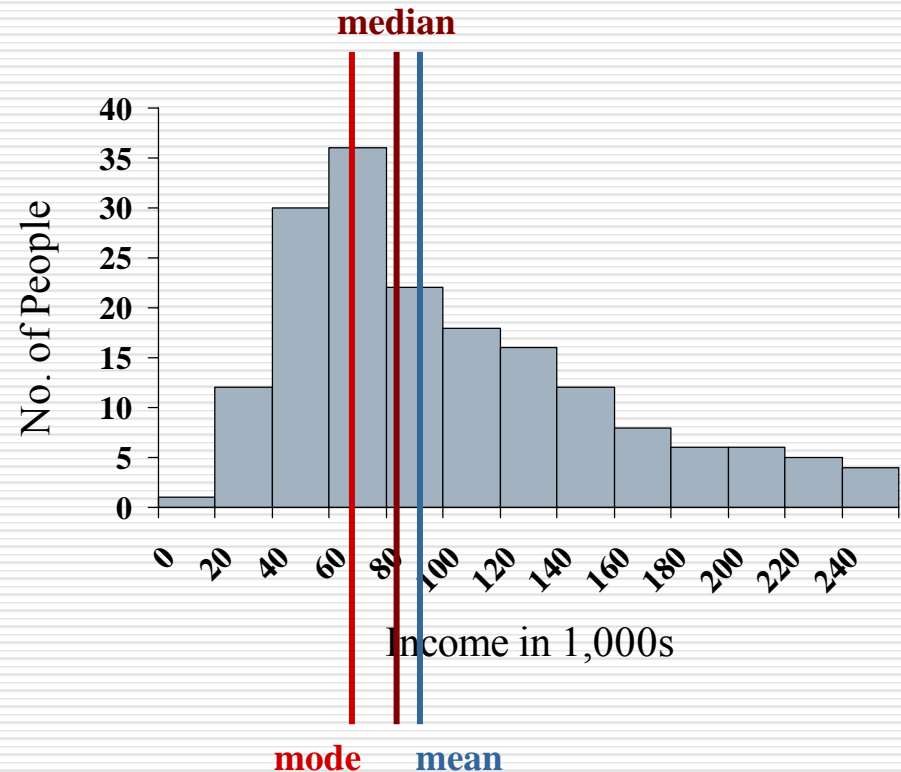
Skewed distributions

- Skewness refers to the asymmetry of the distribution
- A positively skewed distribution is asymmetrical and points in the positive direction.

Mode = 70,000\$

Median = 88,700\$

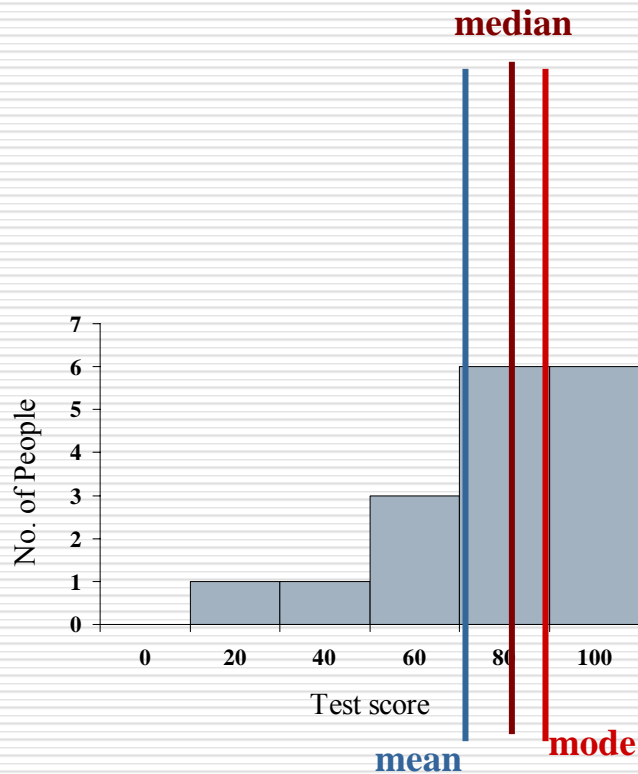
Mean = 93,600\$



• **mode** < **median** < **mean**

Skewed distributions

- A negatively skewed distribution



- **mode** > **median** > **mean**

Skewness and Kurtosis

- ❑ For any symmetrical distribution, the coefficient of skewness (calculated from the variance) = zero
 - ❑ Positive values for coefficient of skewness corresponds to right-skewed distribution
 - ❑ Negative values for coefficient of skewness corresponds to left-skewed distribution
 - ❑ Kurtosis refers to flatness or peakedness of a distribution
 - ❑ Similar to skewness, kurtosis is calculated from variance
 - ❑ For a normal distribution, coefficient of kurtosis = 3
 - ❑ For distributions having more spread than normal (more flat) it is greater than 3 , called platykurtic distribution
 - ❑ For distributions having less spread than normal (more peaked) it is less than 3 , called leptokurtic distribution
-

Using measures of central tendency

2 factors :

The scale of measurement

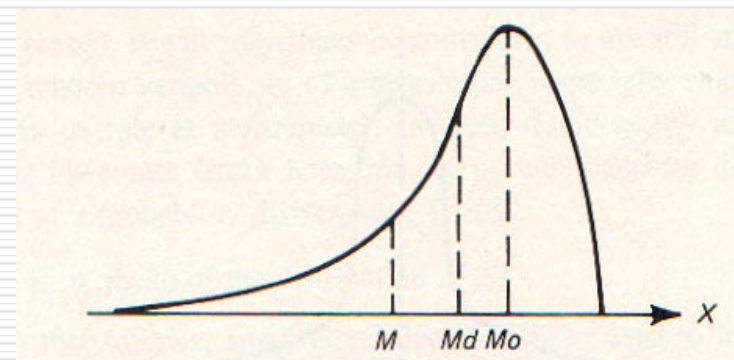
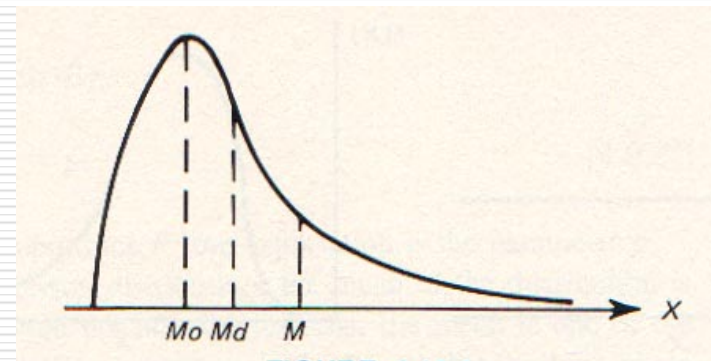
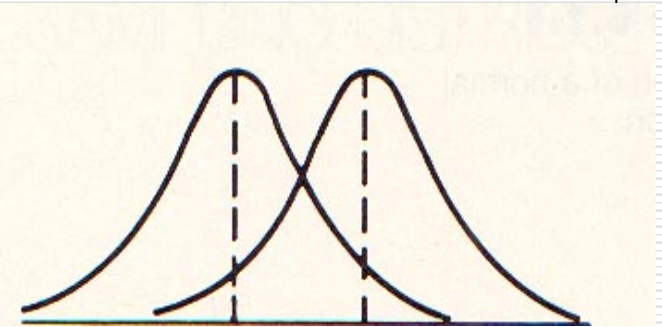
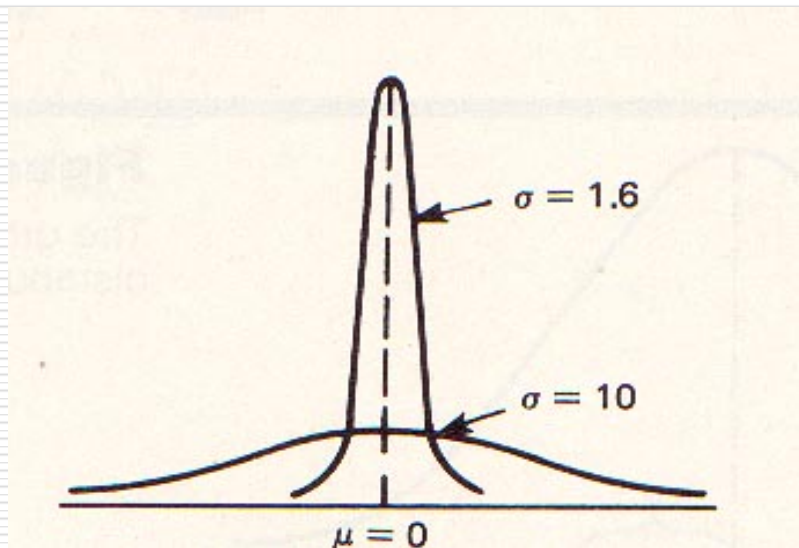
The shape of the distribution

Measures of central tendency

	+	-
Mode	<ul style="list-style-type: none">□ quick & easy to compute□ useful for nominal data	<ul style="list-style-type: none">□ poor sampling stability
Median	<ul style="list-style-type: none">□ not affected by extreme scores□ Useful for ordinal data	<ul style="list-style-type: none">□ somewhat poor sampling stability
Mean	<ul style="list-style-type: none">□ sampling stability□ related to variance	<ul style="list-style-type: none">□ inappropriate for discrete data□ affected by skewed distributions

Distributions

- Center: mode, median, mean
- Shape: symmetrical, skewed
- Spread, variability



Measures of Position

Quantile

Deviation Score

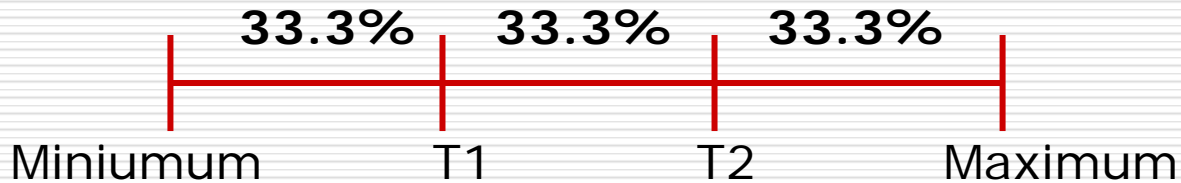
Z-score

Quantiles

Values that divide a data set into equal parts, after ranking data set

1 - Tertile

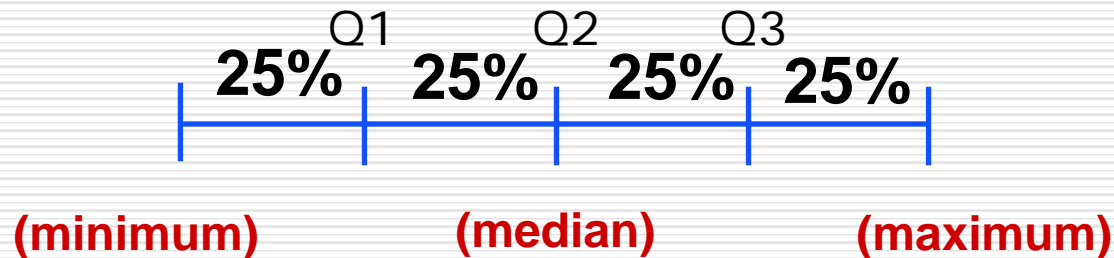
Divides ranked scores into three equal parts



Quantiles

2 - Quartile

Divides ranked scores into four equal parts



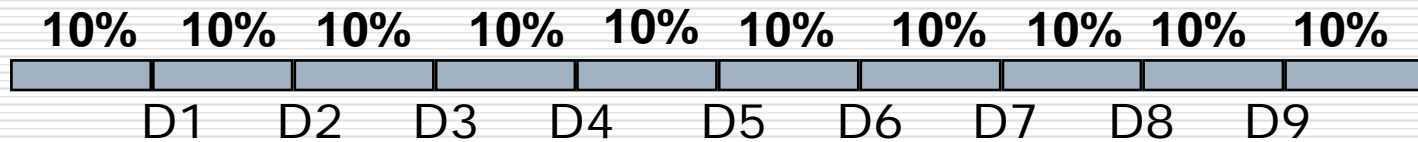
3 - Quintile

Divides ranked scores into five equal parts

Quantiles

4 - Decile

Divides ranked scores into ten equal parts



Quiz:

Median equal to which decile?

Quantiles

5 – Percentile:

Divides ranked scores into 100 equal parts

They are used extensively in medicine to compare individual values with a set of norms

e.g. Physical growth charts

measurements of abilities and intelligence

normal range of laboratory values

normal limits is set between 2.5th and 97.5th percentiles
so, enclosing inbetween 95% of the distribution

Quantiles

For percentiles

- *K^{th} percentile (centile) is the point below which $k\%$ of the values of a distribution lie.*
- *for a distribution with n observations:*

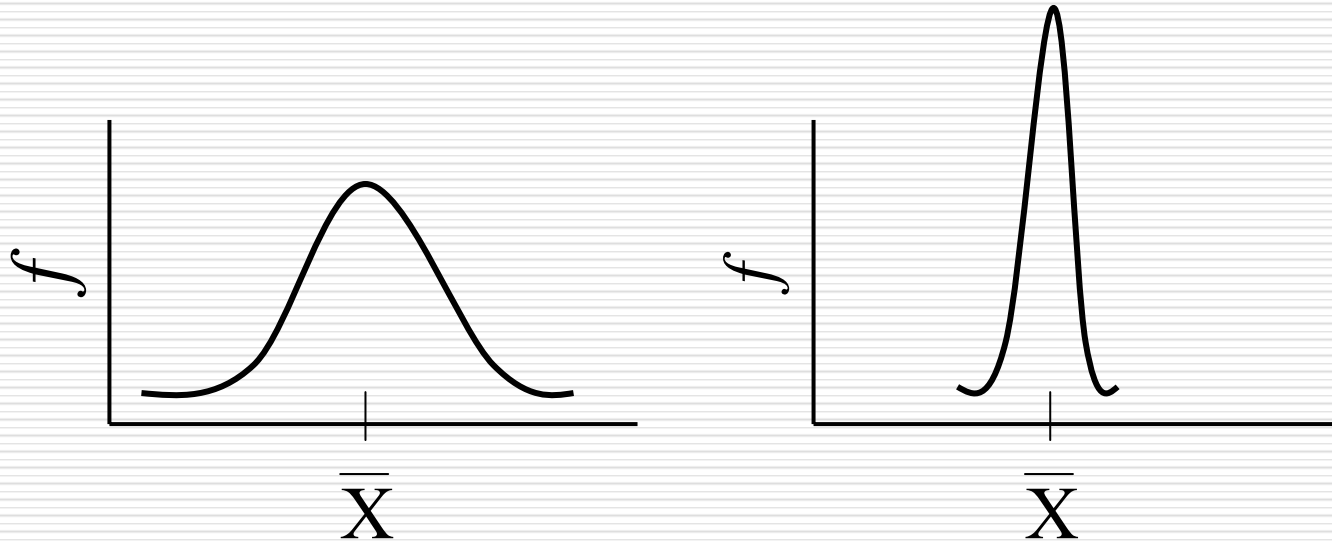
K^{th} percentile = $K (n+1)^{th} / 100$ = value of ordered observation

For quartiles

$Q2 = 2 (n+1) / 4$ = value of ordered observation

Measures of Dispersion and Variability:

Central tendency indicates where scores tend to cluster in a distribution




The extent to which scores are alike or different

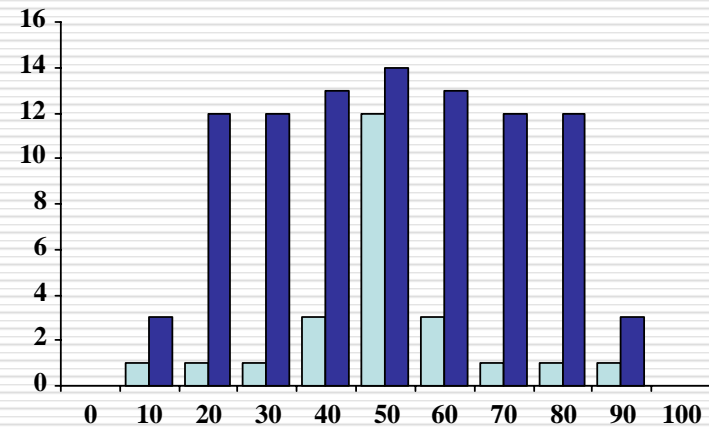
Measures of Spread

- range
 - interquartile range
 - Variance
 - Standard deviation
-

Range

□ largest score minus the smallest score

□ these two  have same range (80)
but spreads look different



□ says nothing about how scores vary around the center

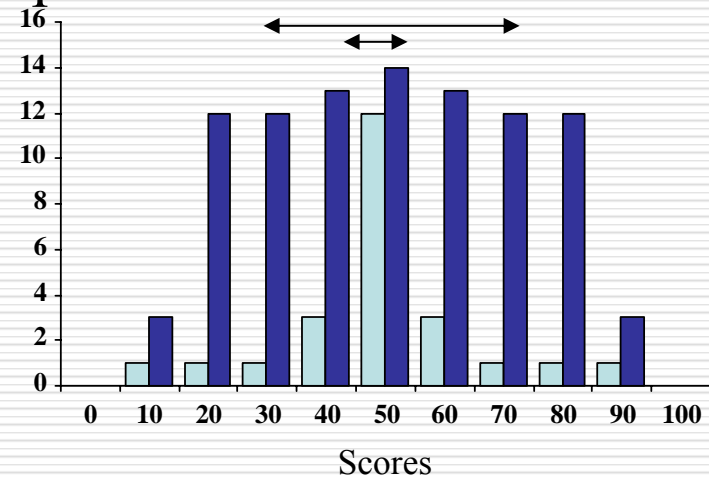
□ greatly affected by extreme scores (defined by them)

Interquartile range

□ the distance between the 25th percentile and the 75th percentile

□ $Q3 - Q1 = 70 - 30 = 40$

□ $Q3 - Q1 = 52.5 - 47.5 = 5$



□ effectively ignores the top and bottom quarters, so extreme scores are not influential

□ dismisses 50% of the distribution

Deviation measures

- Might be better to see how much scores differ from the center of the distribution -- using distance
- Scores further from the mean have higher deviation scores

	Score	Deviation
Hany	10	-40
Ahmed	20	-30
Mahmoud	30	-20
Hesham	40	-10
Liala	50	0
Shereen	60	10
Nabil	70	20
Salah	80	30
Mona	90	40
AVERAGE	50	

Deviation measures

- To see how 'deviant' the distribution is relative to another, we could sum these scores
- But this would leave us with a big fat zero

	Score	Deviation
Hany	10	-40
Ahmed	20	-30
Mahmoud	30	-20
Hesham	40	-10
Liala	50	0
Shereen	60	10
Nabil	70	20
Salah	80	30
Mona	90	40
SUM		0

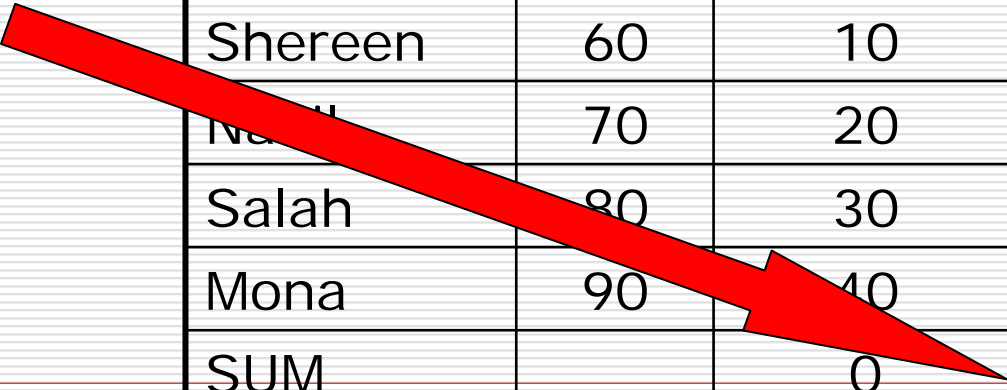
Deviation measures

So we use squared deviations from the mean

This is the sum of squares (SS)

$$SS = \sum (X - \bar{X})^2$$

	Score	Deviation	Sq. Deviation
Hany	10	-40	1600
Ahmed	20	-30	900
Mahmoud	30	-20	400
Hesham	40	-10	100
Liala	50	0	0
Shereen	60	10	100
Nawal	70	20	400
Salah	80	30	900
Mona	90	40	1600
SUM		0	6000



Variance

We take the
"average"
squared
deviation from
the mean and
call it

VARIANCE

For a population:

$$\sigma^2 = \frac{SS}{N}$$

For a sample:

$$s^2 = \frac{SS}{n-1}$$

(to correct for the fact that
sample variance tends to
underestimate pop variance)

Variance

1. Find the mean.
2. Subtract the mean from every score.
3. Square the deviations.
4. Sum the squared deviations.
5. Divide the SS by N or N-1.

	Score	Dev'n	Sq. Dev.	
Hany	10	-40	1600	
Ahmed	20	-30	900	
Mahmoud	30	-20	400	
Hesham	40	-10	100	
Liala	50	0	0	
Shereen	60	10	100	
Nabil	70	20	400	
Salah	80	30	900	
Mona	90	40	1600	
SUM		0	6000	$6000/8=750$

Standard deviation

The standard deviation is the square root of the variance

$$s = \sqrt{s^2} = \sqrt{\frac{SS}{n-1}}$$

The standard deviation measures spread in the original units of measurement, while the variance does so in units squared.

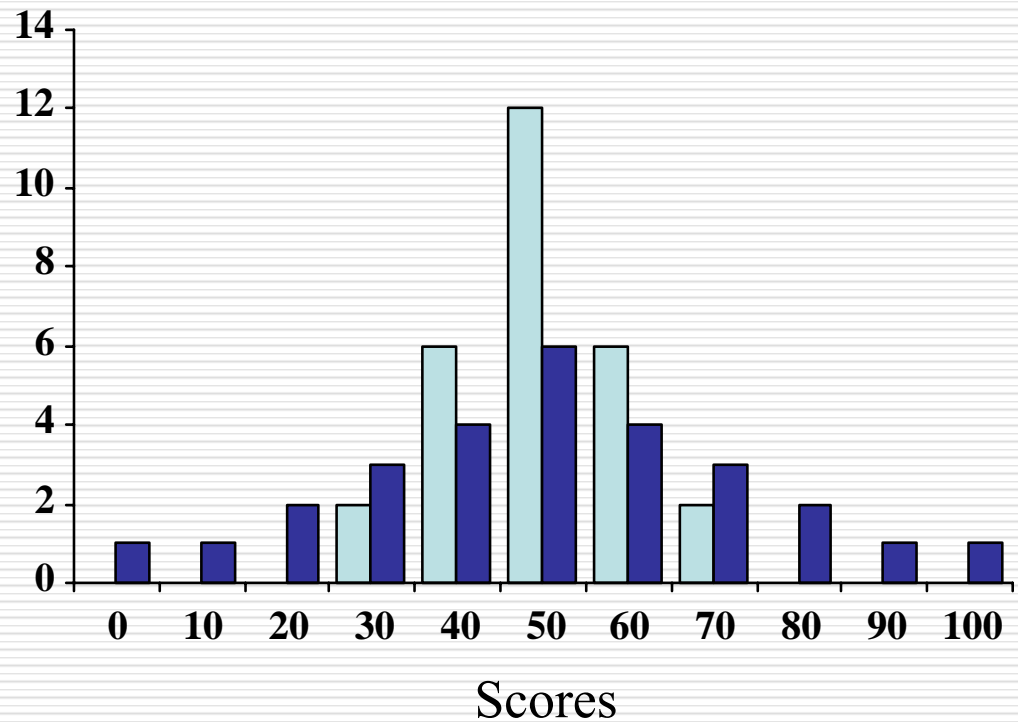
Variance is good for inferential stats.

Standard deviation is nice for descriptive stats.

Example

$N = 28$
 $\bar{X} = 50$
 $s^2 = 140.74$
 $s = 11.86$

$N = 28$
 $\bar{X} = 50$
 $s^2 = 555.55$
 $s = 23.57$



Coefficient of Variation (V or sometimes CV):

Variance (s^2) and standard deviation (s) have magnitudes that are dependent on the magn. of the data.

The coefficient of variation is a relative measure, so variability of different sets of data may be compared (stdev relative to the mean)

$$V = \frac{s}{\bar{X}}$$

Note that there are no units – emphasizes that it is a relative measure
Sometimes expressed as a %

Example:

$$2 = X_1$$

$$2 = X_2$$

$$3 = X_3$$

$$4 = X_4$$

$$5 = X_5$$

$$\overline{X} = 3.2 \text{ g}$$

$$s = 1.304 \text{ g}$$

$$V = \frac{s}{\overline{X}}$$

$$V = \frac{1.304 \text{ g}}{3.2 \text{ g}}$$

$$V = 0.4075$$

or

$$V = 40.75\%$$

Coefficient of variation (MPH)

Sex	N	Mean	Median	StDev	SE Mean		
female	126	91.23	90.00	11.32	1.01		
male	100	129.79	110.00	14.39	1.74		
		Minimum	Maximum	Q1	Q3		
female		65.00	120.00	85.00	98.25		
male		75.00	162.00	95.00	118.75		

Females: $CV = (11.32/91.23) \times 100 = 12.4$

Males: $CV = (14.39/129.79) \times 100 = 11.1$

$$\text{Sample SS} = \sum (X_i - \bar{X})^2$$

Sample variance

$$S^2 = \frac{SS}{n - 1}$$

Sample standard dev

$$s = \sqrt{s^2}$$

Coefficient of Variation

$$V = \frac{s}{\bar{X}}$$

Descriptive Statistics: Quick Review

	For a population:	For a sample:
Mean	$\mu = \frac{\sum X}{N}$	$\bar{X} = \frac{\sum X}{n}$
Variance	$\sigma^2 = \frac{SS}{N}$	$s^2 = \frac{SS}{n-1}$
Standard Deviation	$\sigma = \sqrt{\sigma^2}$	$s = \sqrt{s^2}$

Other measures of central tendency

Geometric mean GM or G

- It is the n^{th} root of the product of the observations
 - In symbolic form for n observations $X_1, X_2, X_3, \dots, X_i$ the geometric mean is
 - $GM = n^{\text{th}} \sqrt{(X_1)(X_2)(X_3)\dots(X_i)}$
 - *Used with data measured on logarithmic scale*
-

Weighted average

It is often necessary to average means or other statistics (OR, Vital statistics) that may differ in their reliabilities because, for example, they are based on different sample sizes. In such cases a weighted average needs to be computed.

$$\bar{X}_w = \frac{\sum_{i=1}^N w_i X_i}{\sum_{i=1}^N w_i}$$

Example

If the following three means are based on differing sample sizes, their weighted average will be:

$$\bar{X}_w = \frac{12 \cdot 3.85 + 25 \cdot 5.21 + 8 \cdot 4.70}{12 + 25 + 8} = 4.76$$

which differs from the unweighted average

$$\bar{X} = \frac{3.85 + 5.21 + 4.70}{3} = 4.59$$

\bar{X}_i	n_i
3.85	12
5.21	25
4.70	8