Biostatistics

L3

Associated professor. Dr. Nedal Alnawaiseh: M. B. Ch. B (MD), Baghdad, Iraq. MSc, JUST, Jordan. MSPH, Tulane University, USA. PhD, UKM, Malaysia. PhD, UNU, IIGH.
Public Health & Community Medicine Department, Medical School, Mutah University, Jordan. Mobile: +962795891817
nidalnawayseh@yahoo.com

Review

1. INFERENTIAL STATISTICS An area of statistics that is concerned about methods of drawing conclusions about a population based on a sample.

2. A PARAMETER is a piece of numerical information about a POPULATION, and a STATISTIC is a piece of numerical information about a SAMPLE.

N.B. The random variable from which a statistic is calculated is referred to as an ESTIMATOR.

Descriptive Methods

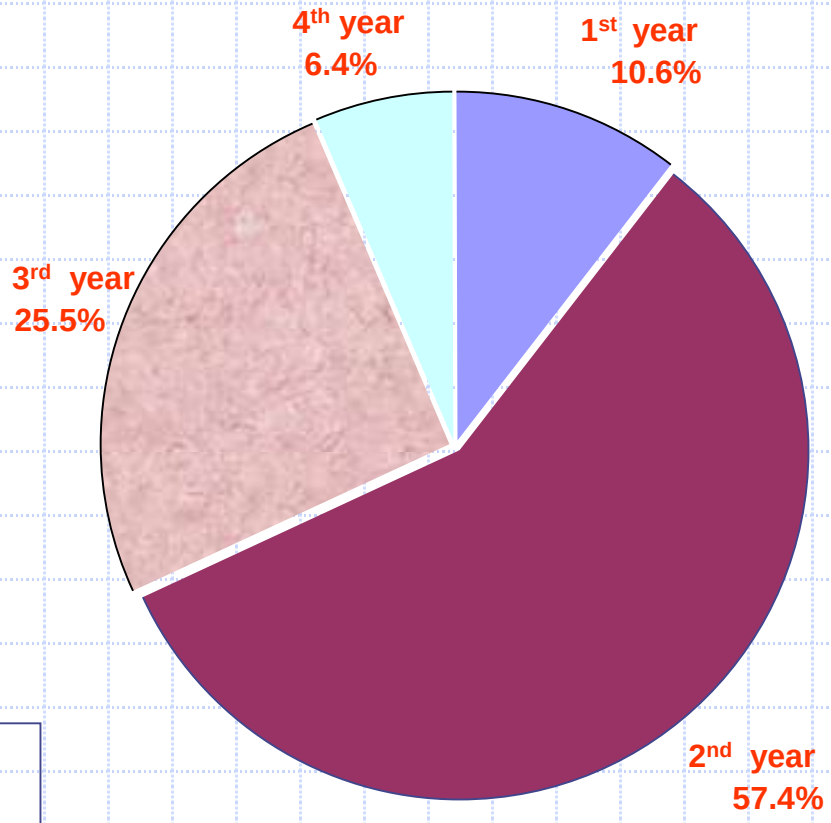
- Diagrams
  - Pie Charts
  - Bar Graphs
  - Histograms
- Measures of central tendency
  - mean
  - median
  - mode
- Measures of dispersion
  - sample variance
  - sample standard deviation

Pie Charts

- Displays data in percentages.
- Statistics Class Data:
  - 5: 1st year, 10.6%
  - 27: 2nd year, 57.4%
  - 12: 3rd year, 25.5%
  - 3: 4th year, 6.4%
- Should add to 100%, adds to 99.9% due to round-off error

Excellent in showing
part vs. whole comparisons

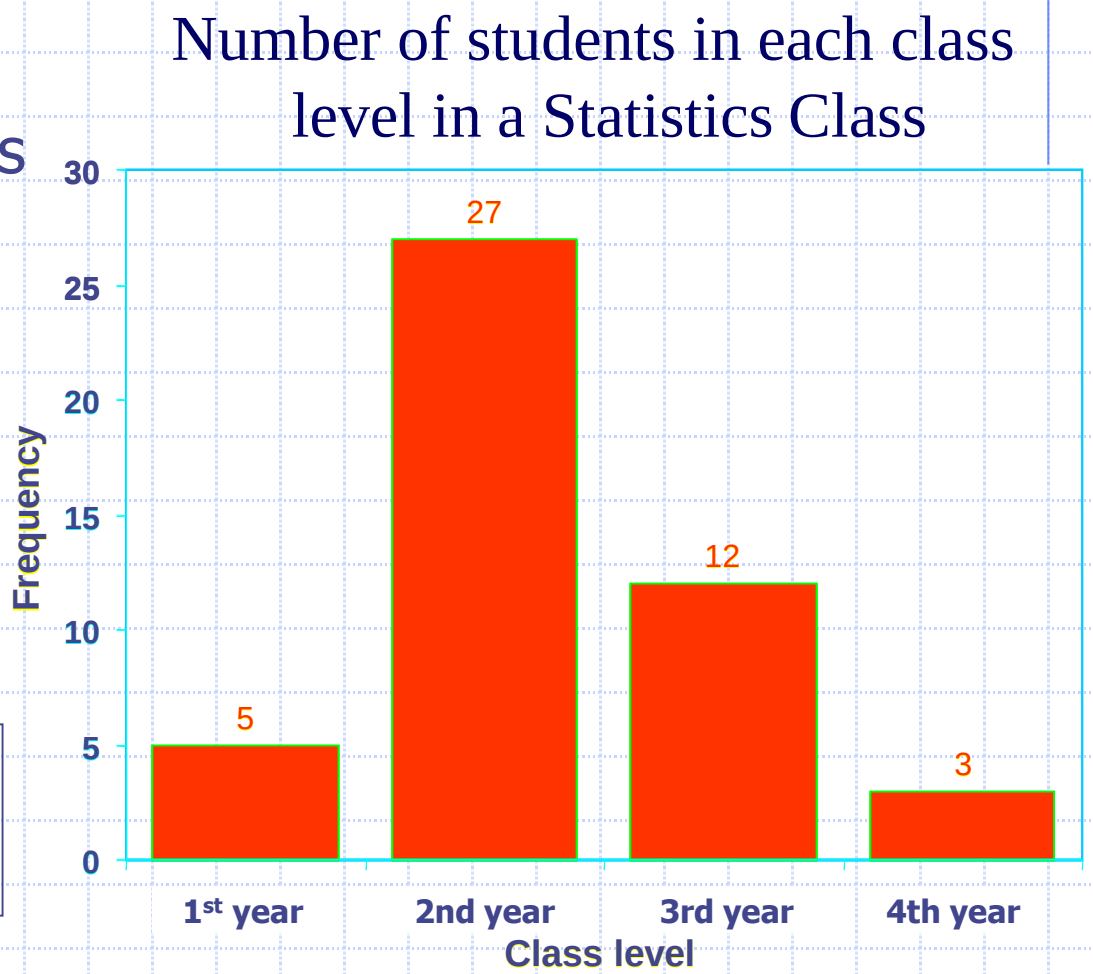
Percentage of students in each class level in a Statistics class



Bar Graphs: Using frequencies

- Example using counts
- Statistics Class Data:
  - 5 1st year
  - 27 2nd year
  - 12 3rd year
  - 3 4th year

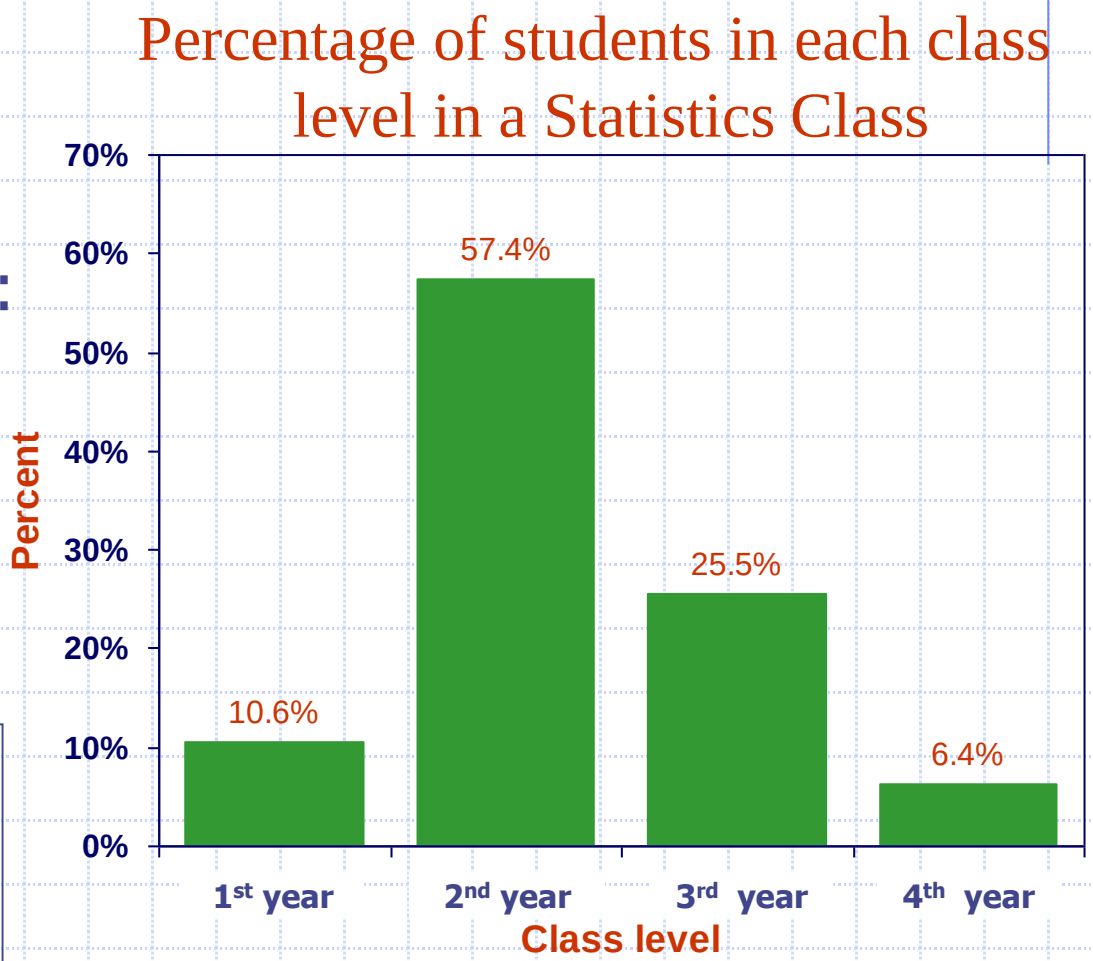
Excellent for showing
Magnitude differences



Bar Graphs: Using Percentages

- Example using percentages
- Statistics Class Data:
  - 10.6% 1st year
  - 57.4% 2nd year
  - 25.5% 3rd year
  - 6.4% 4th year

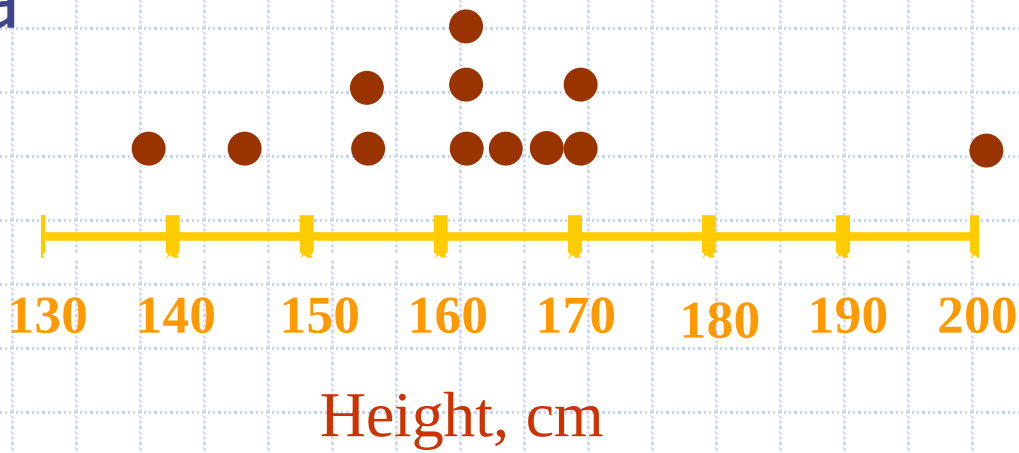
Allows easier comparisons between data sets of different sizes.



Dotplot

- Number line with dots representing data points
- Can visualize the "spread" of the data
- Data: Height of of 12 female students measured in (cm)

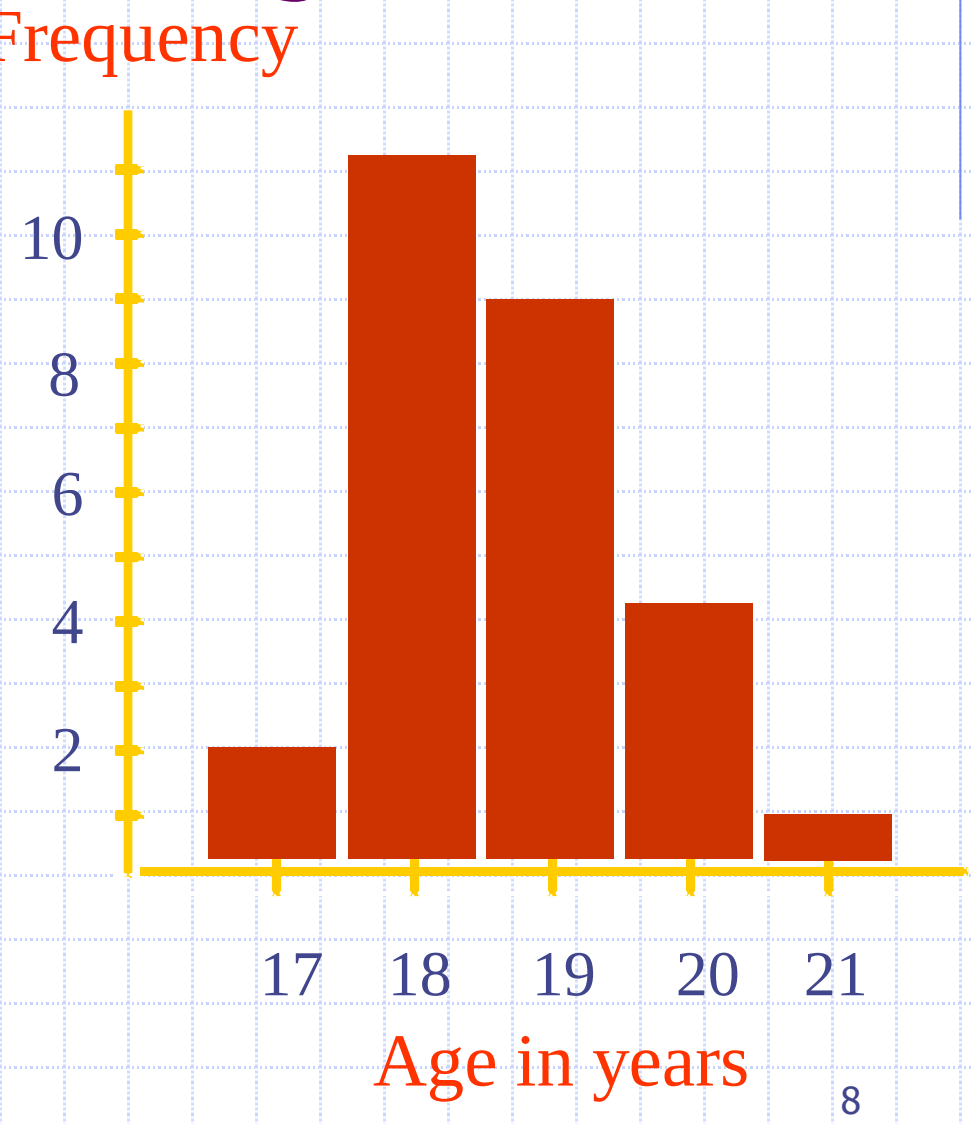
  139, 161, 170, 201, 161, 168, 170, 155, 165, 145, 155, 161



Ungrouped Frequency Distribution and Histogram

Ages of students in a statistics class

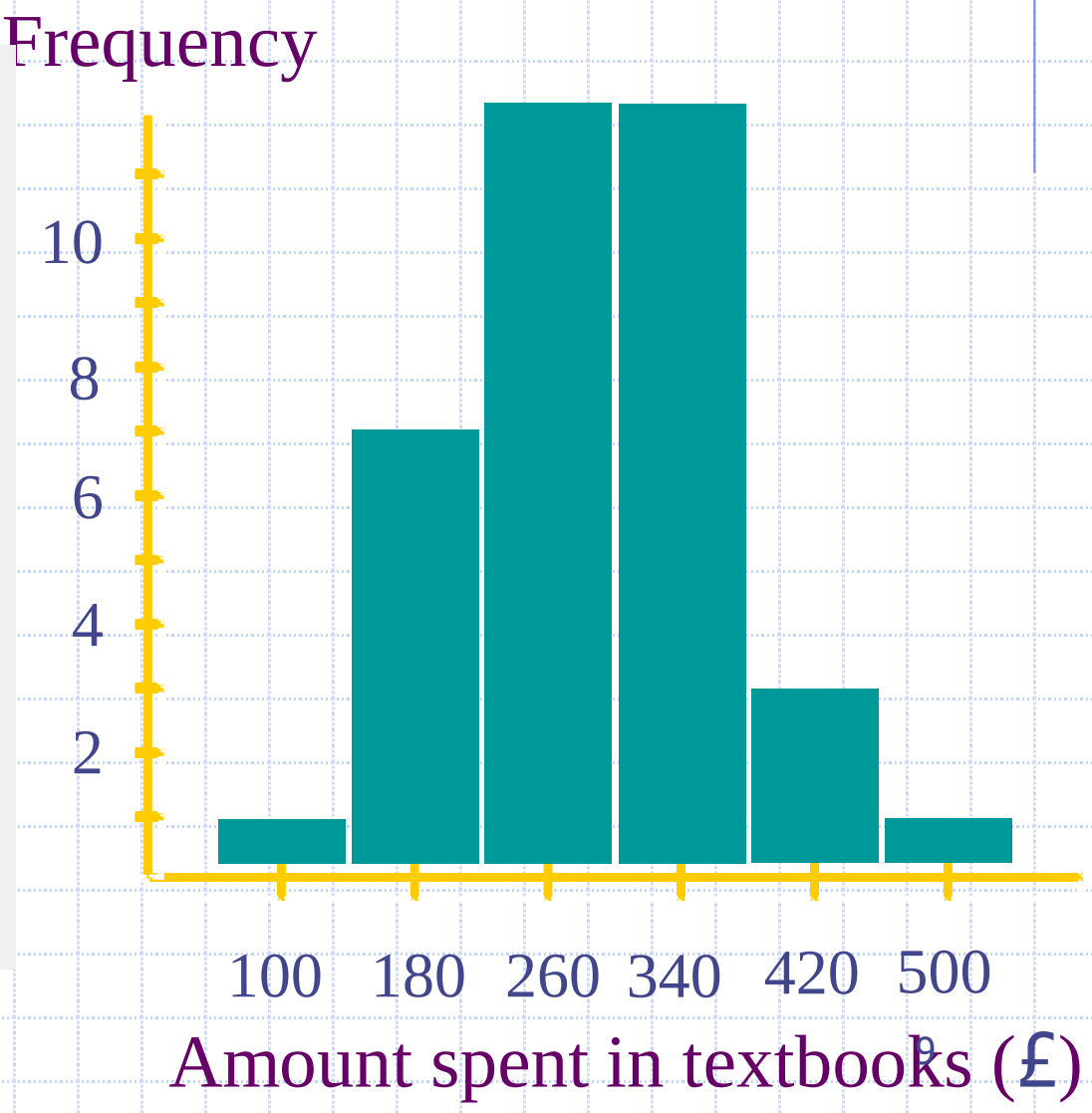
Age	Frequency
17	2
18	11
19	9
20	4
21	1



Grouped Frequency Distribution and Histogram

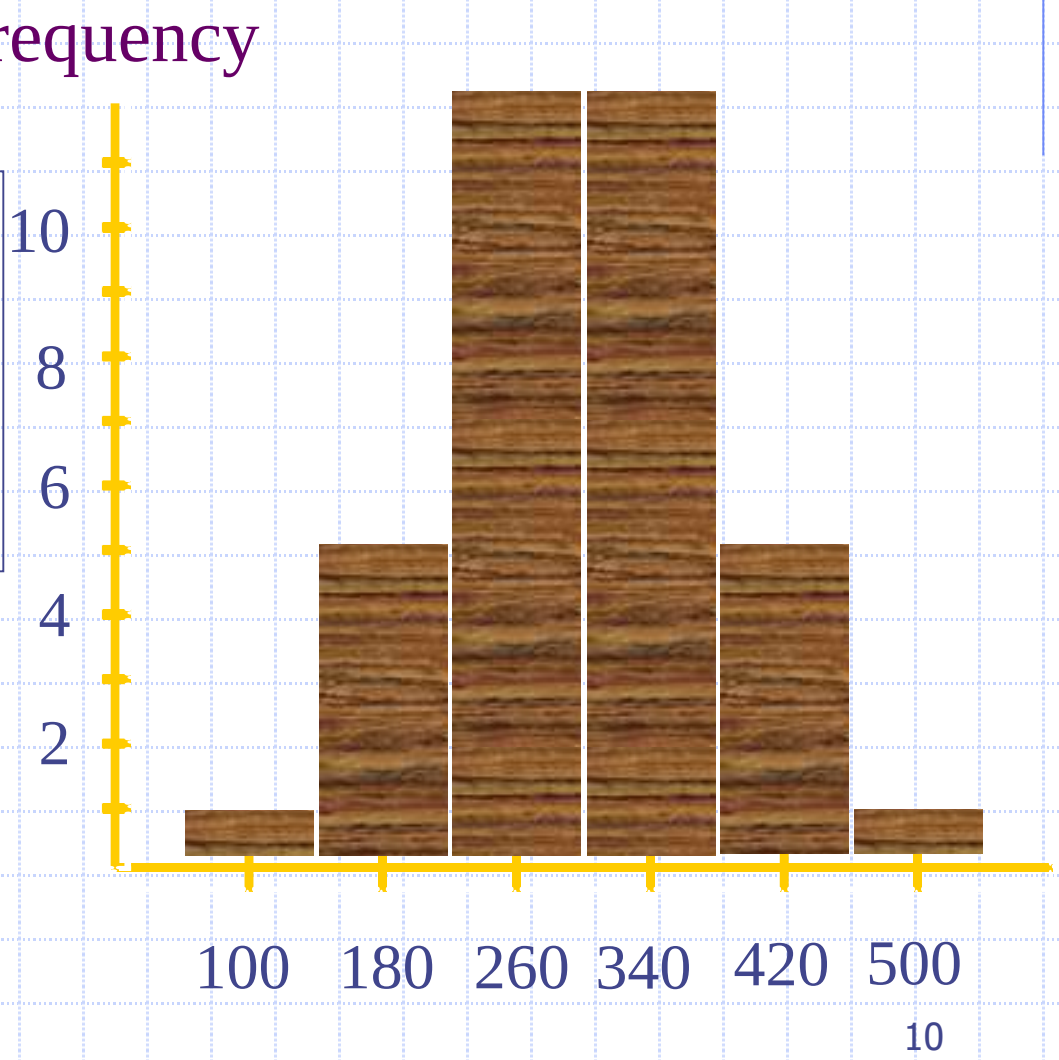
Amount spent on textbooks per student:

Amount (£)	Frequency
60-139	1
140-219	7
220-299	12
300-379	12
380-459	3
460-539	1



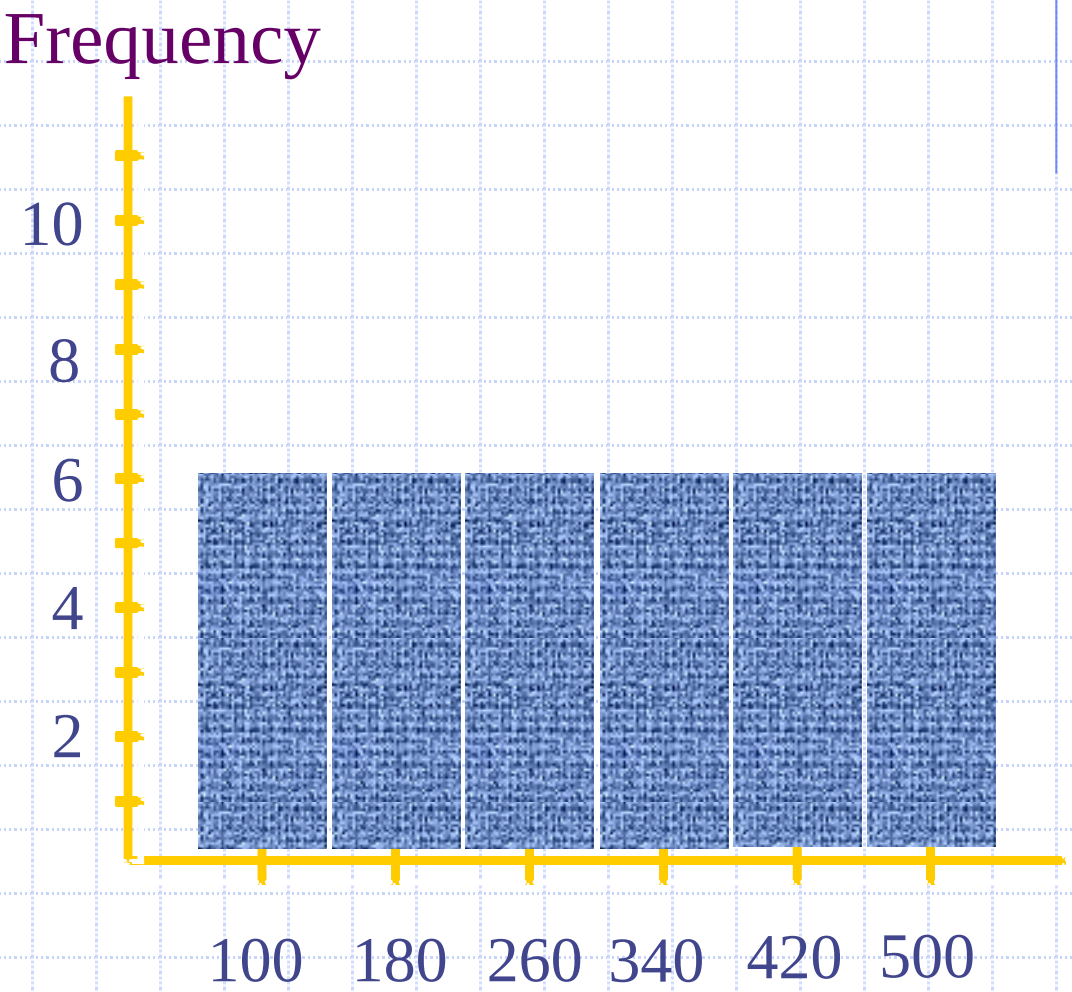
Shapes of Histograms I

Symmetrical,
normal,
or bell-shaped



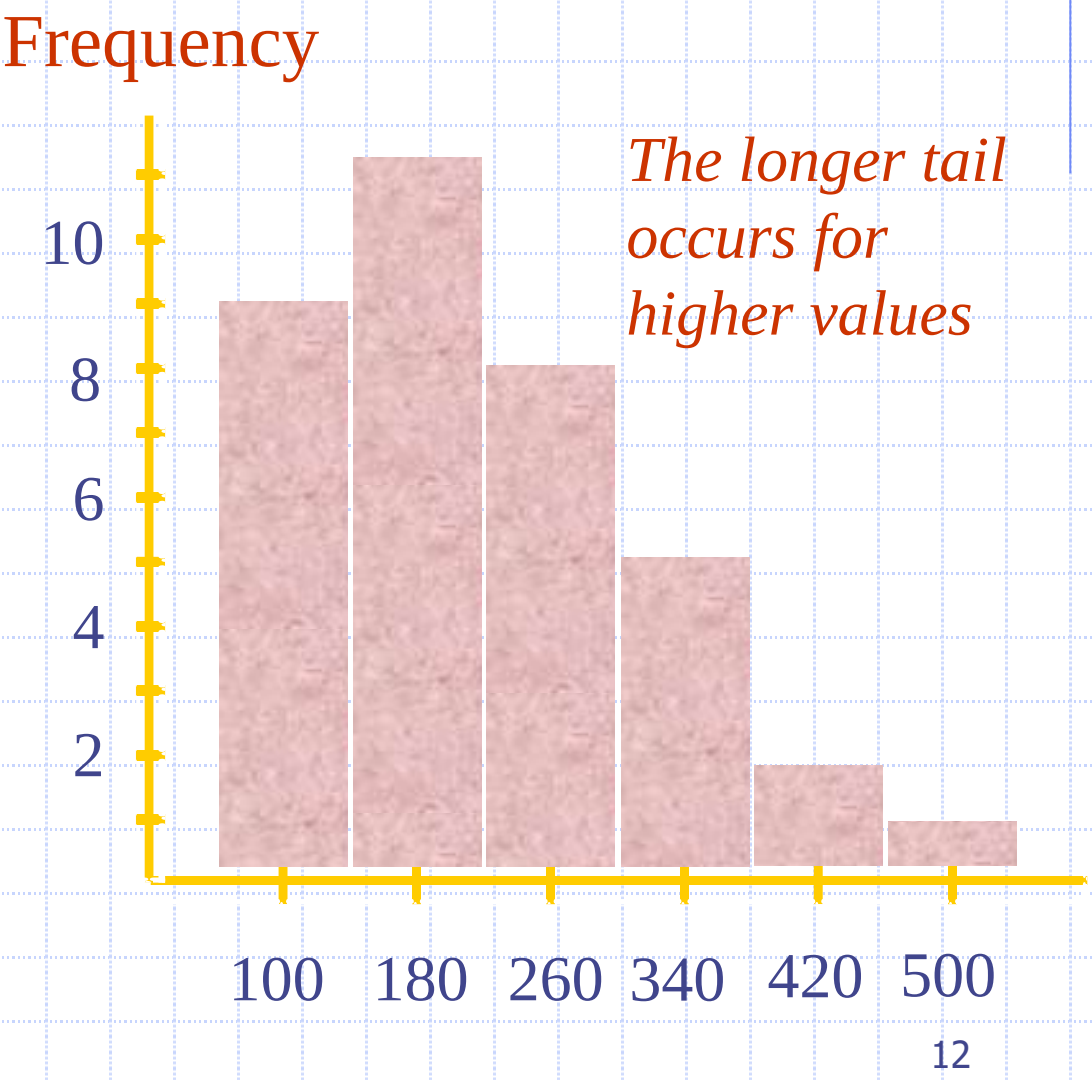
Shapes of Histograms II

Uniform or rectangular



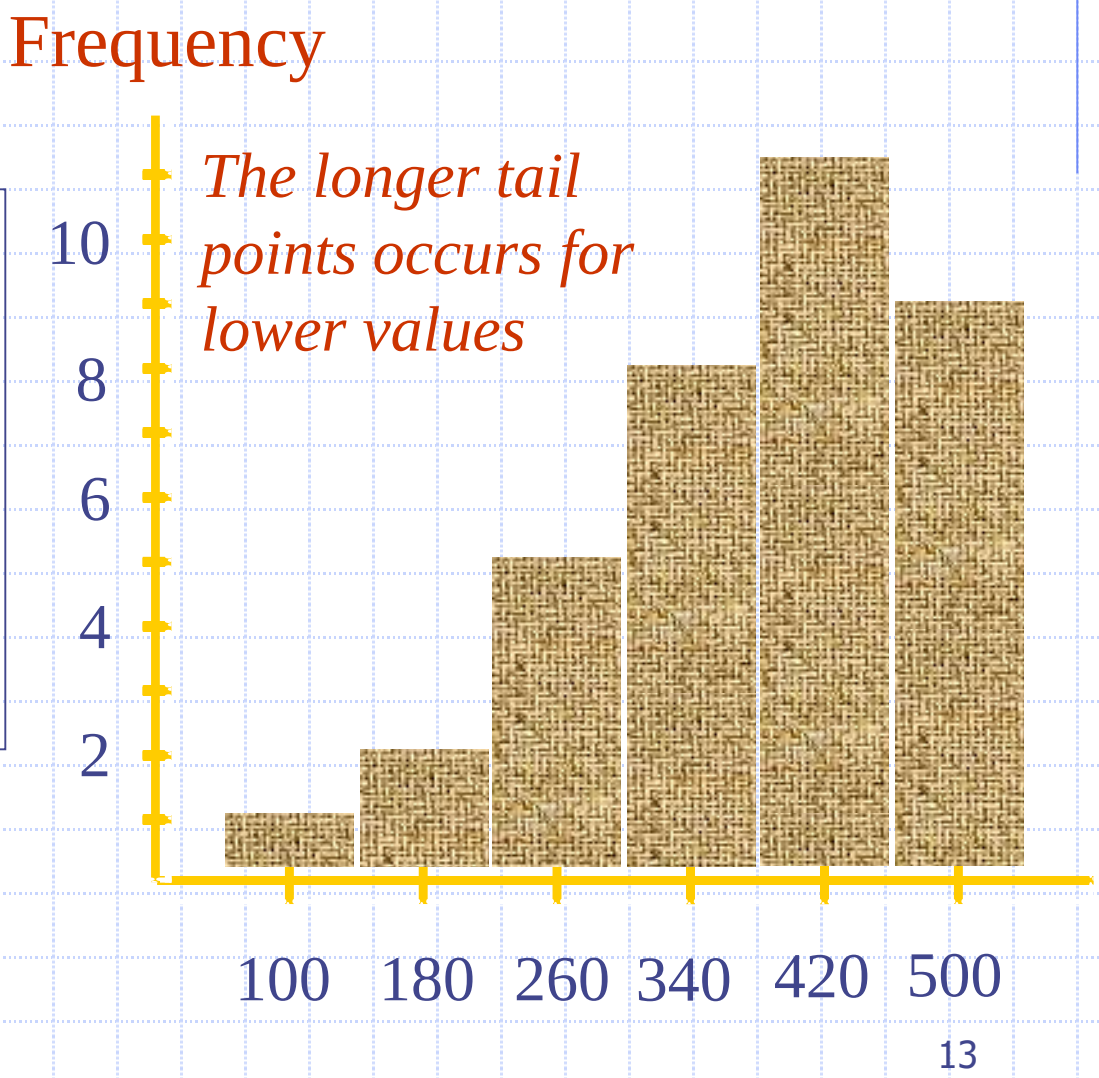
Shapes of Histograms III

Skewed right
or
Positively
skewed



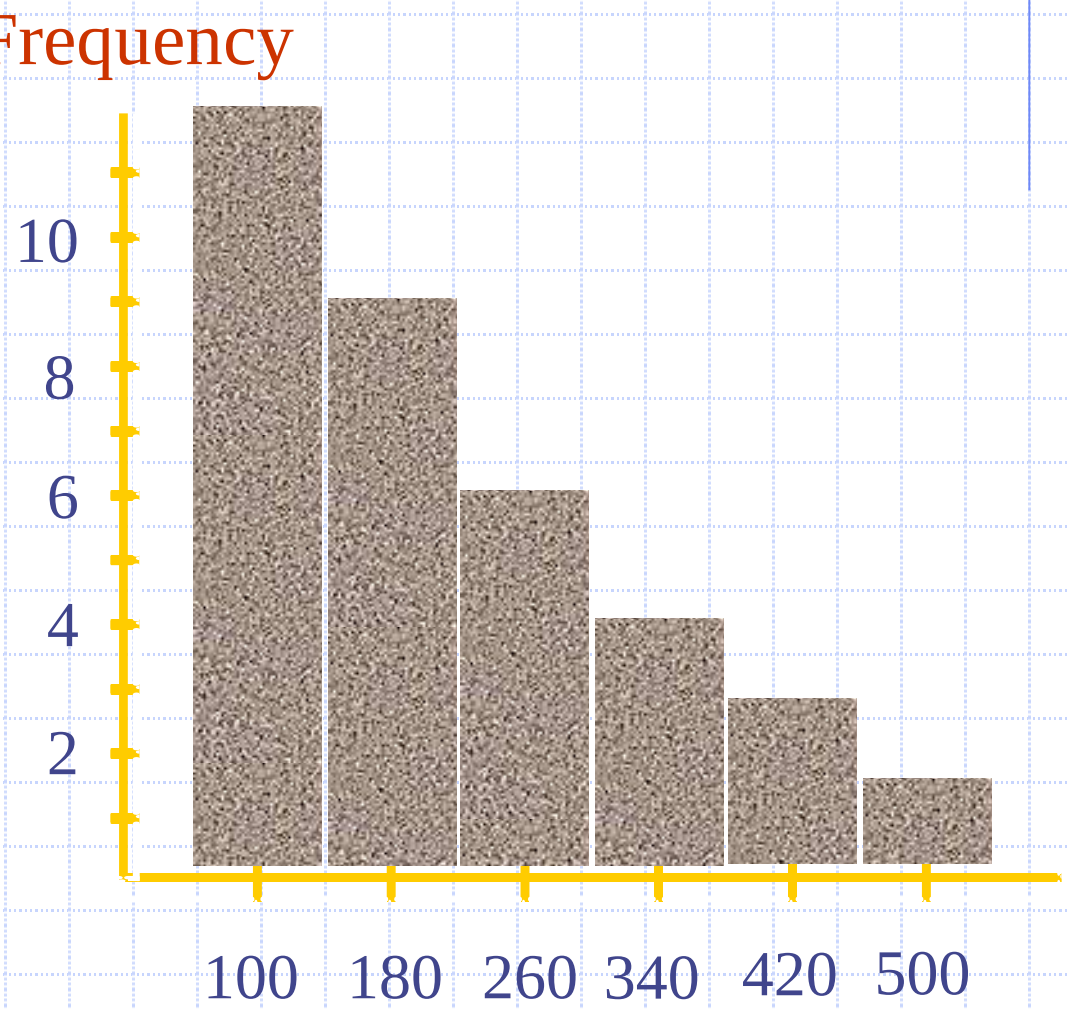
Shapes of Histograms IV

Skewed left
or
Negatively
skewed



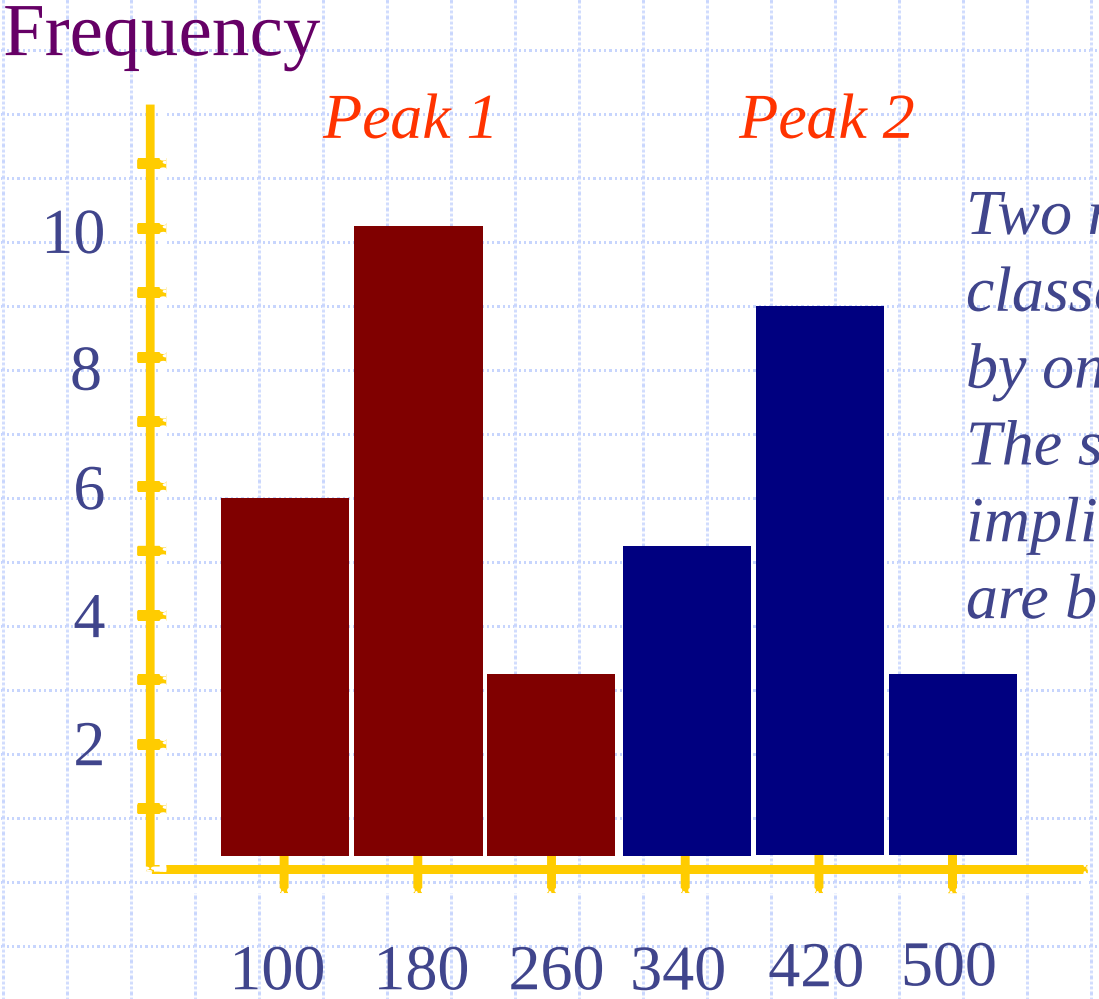
Shapes of Histograms V

J-shaped



Shapes of Histograms VI

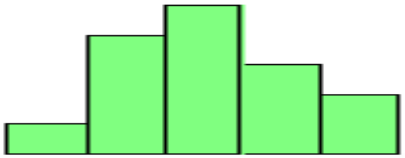
Bimodal



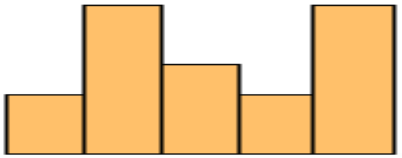
Two most populous classes are separated by one or more classes. The situation often implies that 2 populations are being sampled.

SHAPES OF DISTRIBUTIONS

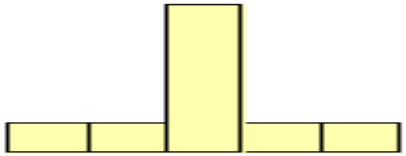
Unimodal



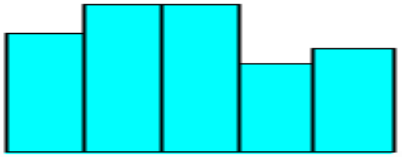
Bimodal



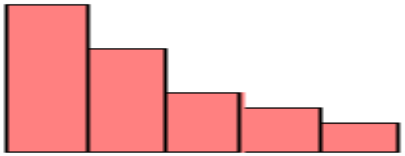
Small Variability



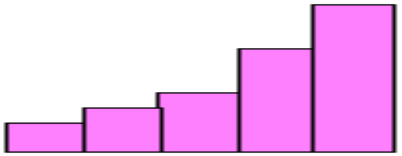
Large Variability



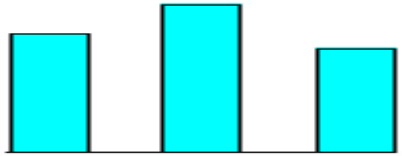
Positively Skewed



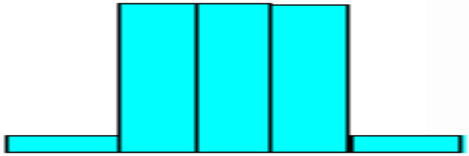
Negatively Skewed



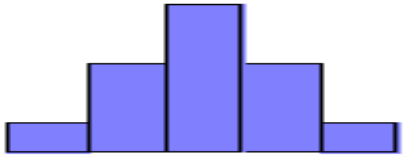
Large Kurtosis

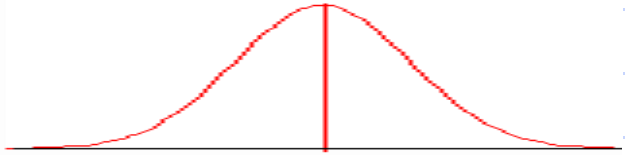


Little Kurtosis



Symmetrical and possibly Normal





Measures of Central tendency

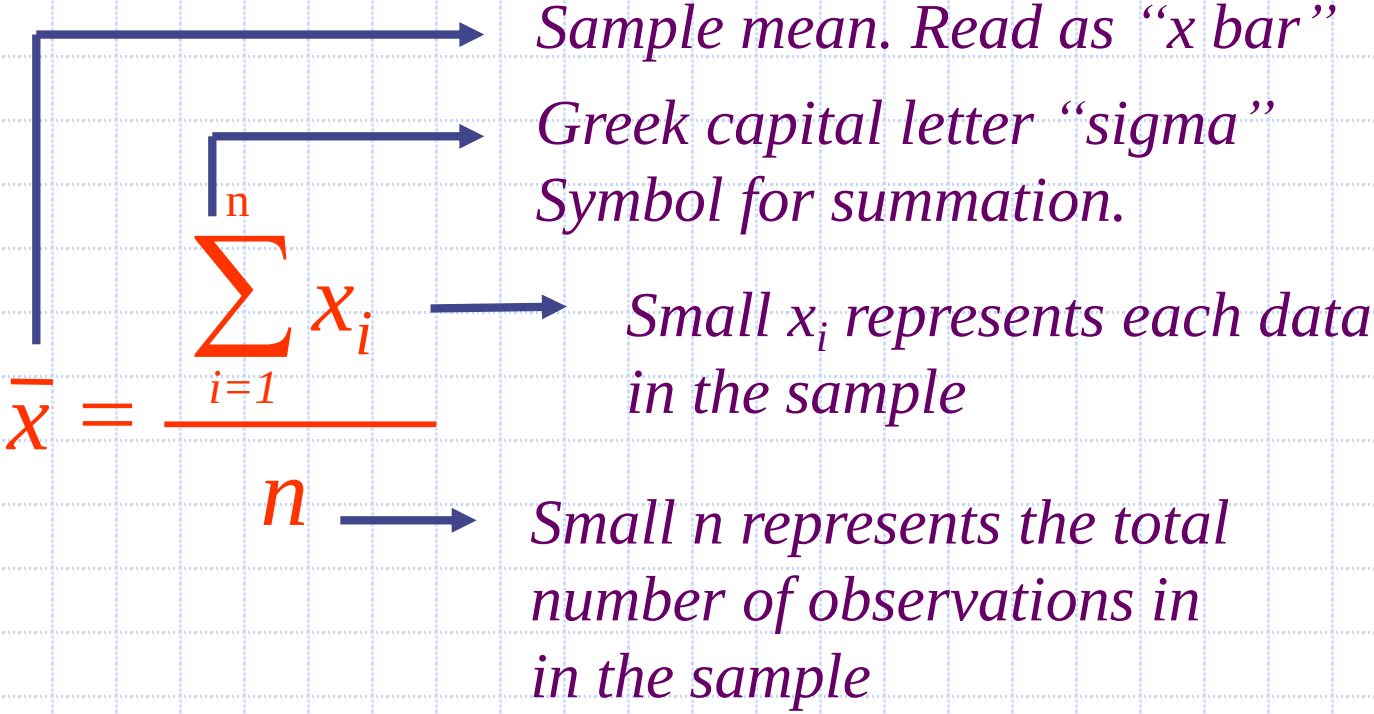
If we would like to give an indication of a "typical", or "most likely" value in a sample, we need to choose a measure of central tendency.

Commonly-used measures of central tendency are:

1. mean
2. median
3. mode

Sample Mean

Sum of the observation values divided by the sample size or the number of observations.



This is an ESTIMATOR. The value calculated from a particular sample is a STATISTIC.

EXAMPLE: Data for the number of faulty items produced per day on a production line: 6, 3, 8, 6, 4.

Find the sample mean. Solution:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{6+3+8+6+4}{5} = \frac{27}{5} = 5.4$$

What would happen if data point 5 were 40? Then the mean would increase significantly ($\bar{x} = 12.6$).

The mean is sensitive to extreme values.

Sample Median

The sample median is the middle value when the data is listed in ascending or descending order.

To get the median we must do the following:

1. Rank data from least to greatest
2. If the number of observations (n) is odd, then the median ($\tilde{x}$) is the centre value. If n is even, then the median is the mean of two middle values

Example I

Number of faulty items 6, 3, 8, 6, 4.
Find the Sample Median.

Step 1: Rank data (3, 4, 6, 6, 8)

Step 2: Since n =5 is odd, then the median is the 3rd value

Hence $\tilde{x} = 6$

Example II

Compute the median of the number of children in eight families (2,3,1,4,3,2,3,3)

Step 1: Rank data (1, 2, 2, 3, 3, 3, 3, 4)

Step 2: Since n =8 is odd, then the median is the mean of the 4^{th} and 5^{th} value

$$\tilde{x} = \frac{3+3}{2} = 3$$

Properties of the Median

- The median separates the ranked set of data into two equal parts, by which we mean that 50% of the observations are below the median and 50% are above the median.

- The median is not as sensitive to extreme values as the mean.

Sample Mode

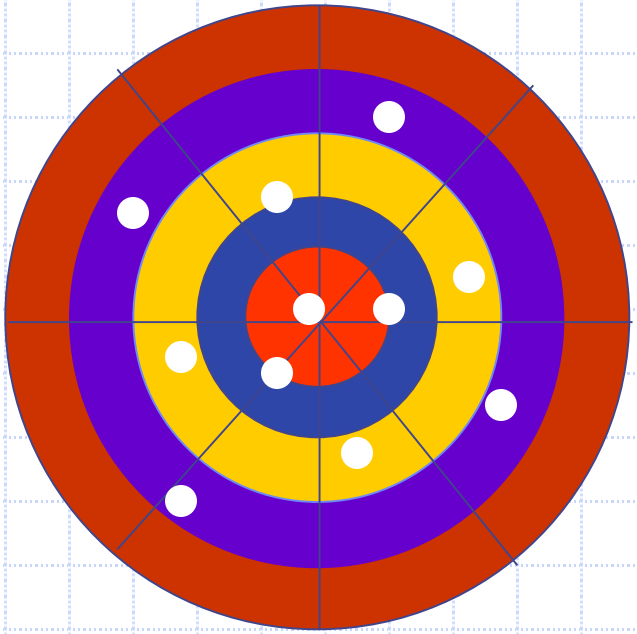
Most frequently occurring value

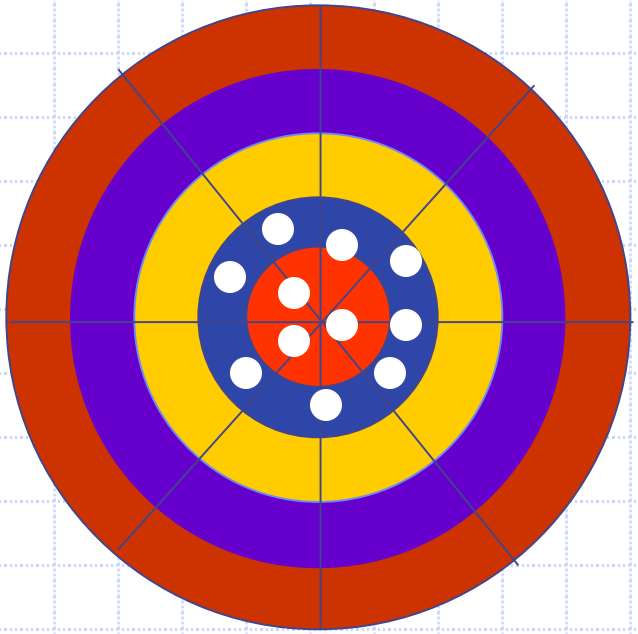
- This is the only measure of central tendency that can be used for qualitative data
- Does not always exist

Example: Recall children data (1, 2, 2, 3, 3, 3, 3, 4)

Mode = 3

Measures of Dispersion





SHOOTER A SHOOTER B

Both shooters are hitting around the "centre" but shooter B is more "accurate"

Deviation from the mean

The sum of the differences between any data value and the sample mean:

$$\sum_{i=1}^{n} (x_i - \bar{x})$$

Seems to be a possible candidate for a measure of dispersion…

It is not…

Recall data: 6, 3, 8, 6, 4 and $\bar{x} = 5.4$)

x	$x - \bar{x}$
6	0.6
3	-2.4
8	2.6
6	0.6
4	-1.4
Total, $\sum (x - \bar{x})$	0

- Since the sum of the deviations from the sample mean is *always zero*, then this is not a useful statistic

- An alternative would be to take the sum of the squared deviations, which are always positive.

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

Sample Variance

The mean of the squared deviations calculated using n-1 as a divisor

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

Example

Recall data: 6, 3, 8, 6, 4 and $\bar{x} = 5.4$)

x	$x - \bar{x}$	$(x - \bar{x})^2$
6	0.6	0.36
3	-2.4	5.76
8	2.6	6.76
6	0.6	0.36
4	-1.4	1.96
Total (Σ)	0	15.20

The sample variance is:

$$\frac{\sum (x - \bar{x})^2}{n-1} = \frac{15.2}{5-1}$$

$$= \frac{15.2}{4}$$

$$= 3.8$$

An alternative definition

Often the definition of sample variance from the last slide is not easy to use in a pocket calculator. An alternative way is to express this formula in terms of the built-in functions $\sum x$ and $\sum x^2$.

$$s^2 = \frac{1}{n-1}\sum x_i^2 - \frac{n}{n-1}\overline{x}^2$$

Sample Standard Deviation

Sample standard deviation is the positive square root of the variance.

Example: Number of faulty items per day: (6, 3, 8, 6, 4), n =5

The mean is: $\bar{x} = 5.4$

The variance is: $s^2 = 3.8$

The standard deviation is: $s = \sqrt{s^2}$

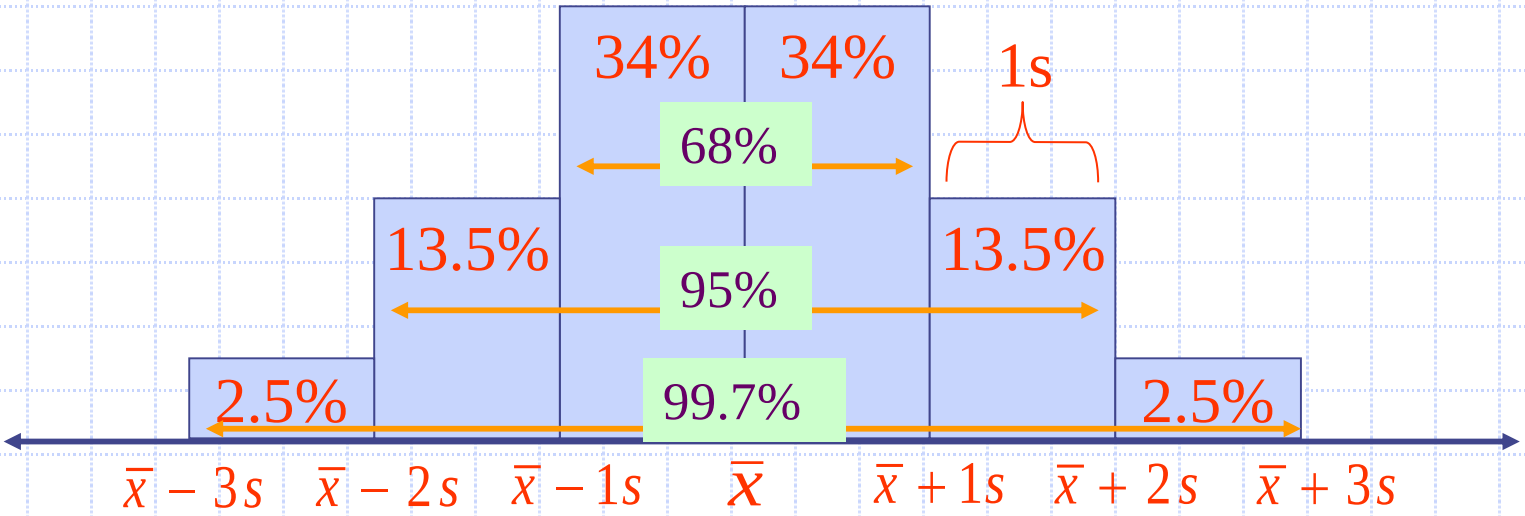
$= \sqrt{3.8}$

$= 1.94935887$

~ 1.9

"most days there are 5.4 ± 1.9 faulty items"

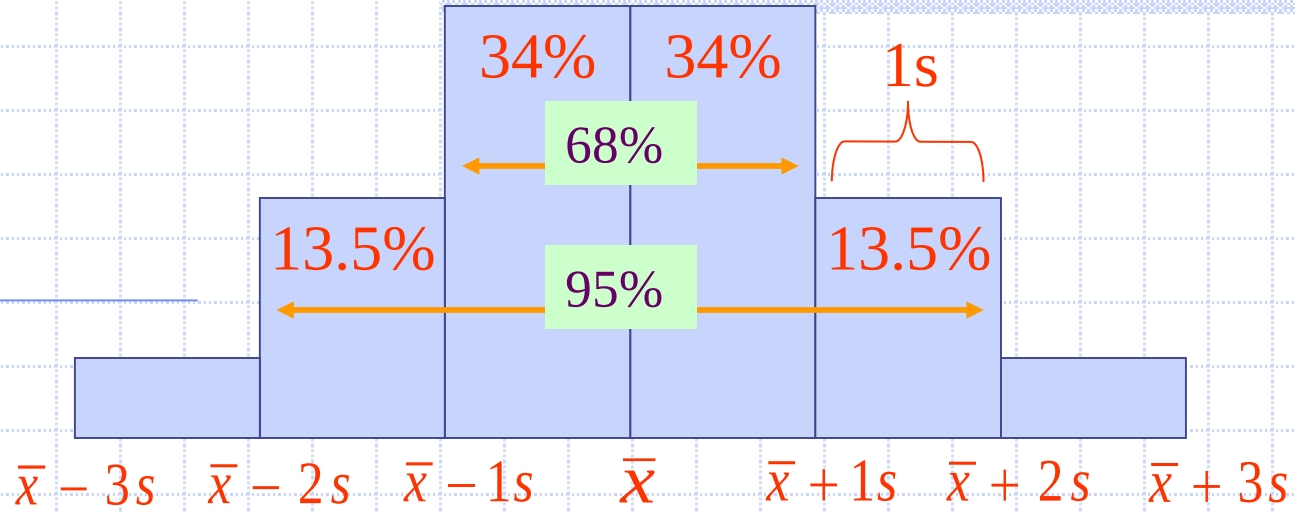
Interpreting Standard Deviation



For bell-shaped shaped distributions, the following statements hold:

- Approximately 68% of the data fall between $\bar{x}-1s$ and $\bar{x}+1s$
- Approximately 95% of the data fall between $\bar{x}-2s$ and $\bar{x}+2s$
- Approximately 99.7% of the data fall between $\bar{x}-3s$ and $\bar{x}+3s$

For NORMAL distributions, the word 'approximately' may be removed from The above statements.



Example: Suppose the amount of liquid in 12 oz. Pepsi cans has a roughly bell-shaped distribution with a mean of 12 oz. and standard deviation of 0.10 oz.

a) Give the interval of the amount of liquid that approximately 68% of the cans will have

12-0.1 to 12+0.1 = 11.9 to 12.1 oz.

b) Give the interval of the amount of liquid that approximately 95% of the cans will have

12-2(0.1) to 12+2(0.1) = 11.8 to 12.2 oz.