Unit Two
Descriptive Biostatistics

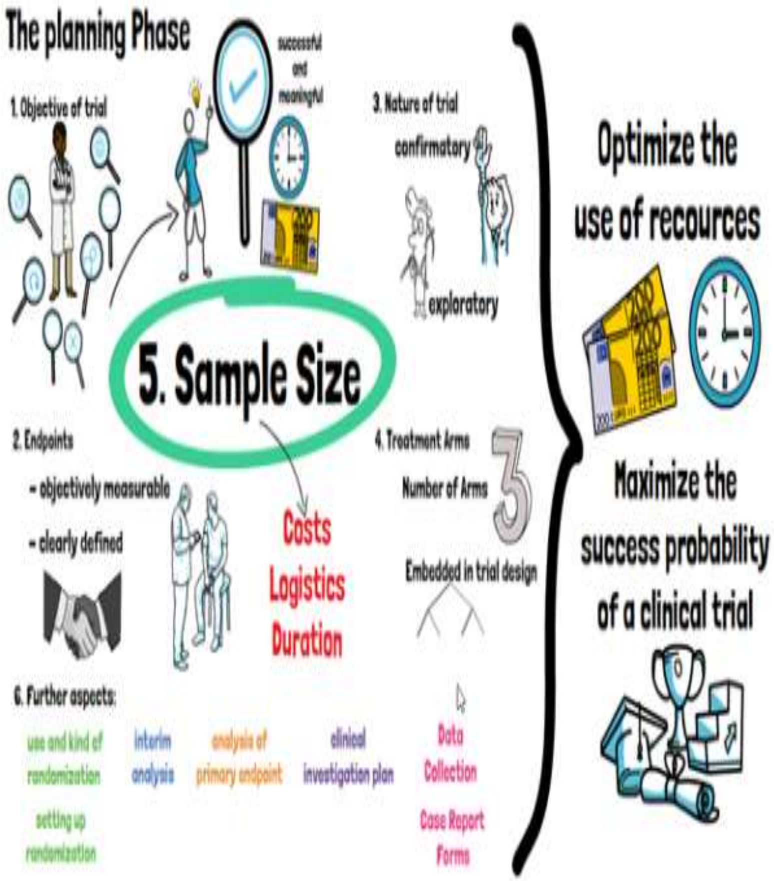
ASSOCIATE PROFESSOR DIANA ARABIAT

Recap

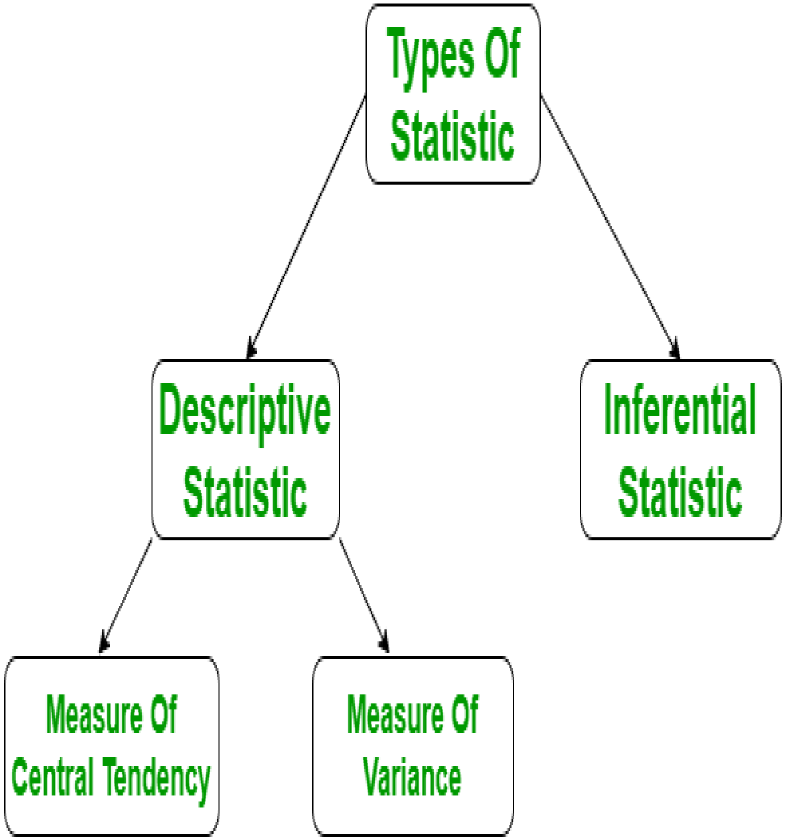
BIOSTATISTICS

What is the biostatistics?

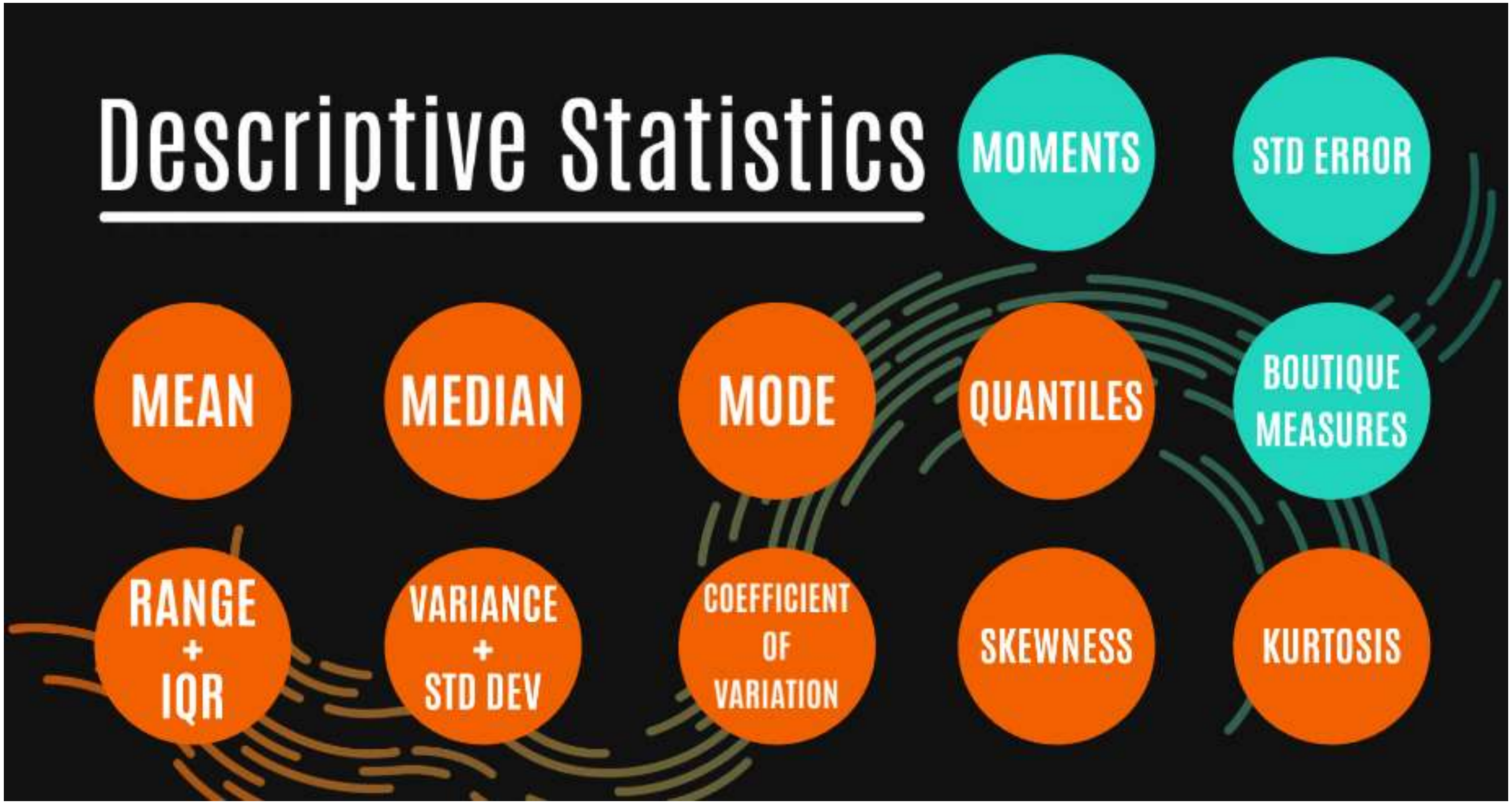
A branch of applied math that deals with collecting, organizing and interpreting data using well-defined procedures.



TYPES OF BIOSTATISTICS:



- Descriptive Statistics. It involves organizing, summarizing & displaying data to make them more understandable.
- Inferential Statistics. It reports the degree of confidence of the sample statistic that predicts the value of the population parameter.


Descriptive Statistics
MOMENTS
STD ERROR
MEAN
MEDIAN
MODE
QUANTILES
BOUTIQUE MEASURES
RANGE + IQR
VARIANCE + STD DEV
COEFFICIENT OF VARIATION
SKEWNESS
KURTOSIS

What is Descriptive Biostatistics?

The best way to work with data is to summarize and organize them.

Numbers that have not been summarized and organized are called **raw data**.

Raw data on number of smarphones owned per family

2	3	4	1	2	2	3	5	2	4
3	2	4	1	3	4	5	3	2	4
2	4	2	3	2	3	2	3	2	3
3	2	3	2	1	2	3	4	1	2
1	2	2	3	3	2	4	2	2	3

Quantitative raw data

Definition

Data is any type of information

Raw data is a data collected as they are received.

Organized data is data organized either in ascending, descending or in a grouped data.

Organizing Data

After collecting data, the first task for a researcher is to organize and simplify the data so that it is possible to get a general overview of the results.

Raw Data: Data which is not organized is called raw data.

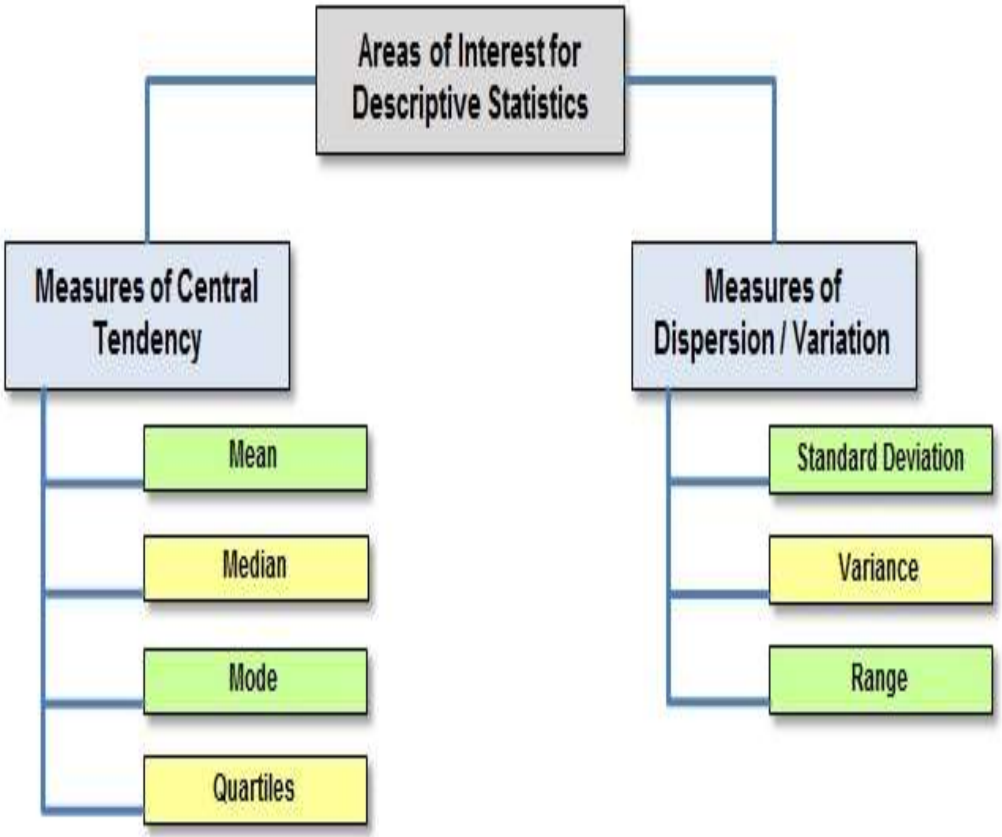
Un-Grouped Data: Data in its original form is called Un-Grouped Data.

Note: Raw data is also called ungrouped data.

Descriptive Measures

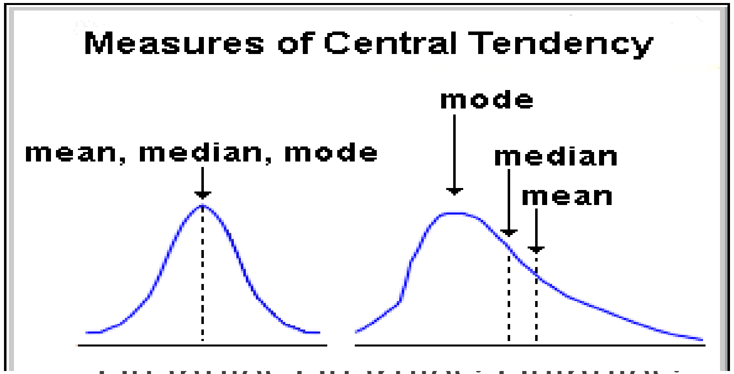
A *descriptive measure* is a single number that is used to describe a set of data.

Descriptive measures include *measures of central tendency* and *measures of dispersion*.



Measures of Location

- Measures of central tendency: Mean; Median; Mode



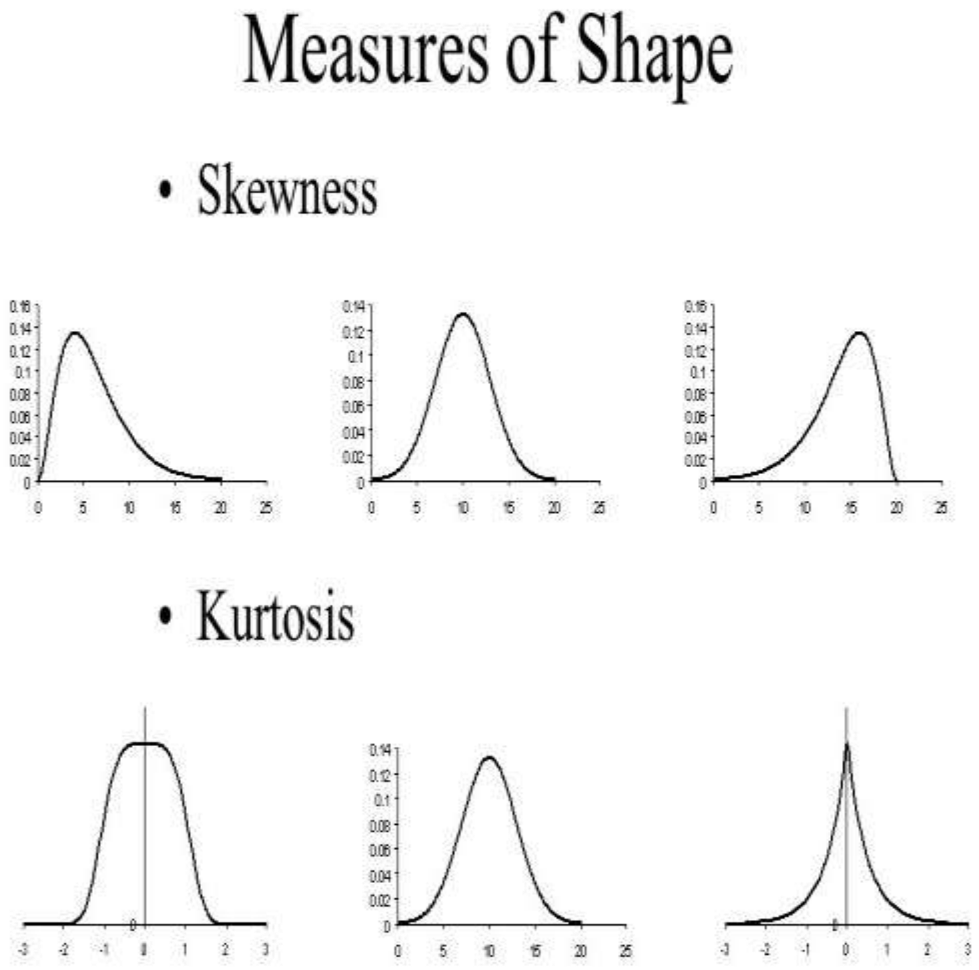
- Quantiles Quartiles; Quintiles; Percentiles

Measures of Location or Central Tendency

This idea of Central tendency refers to the extent to which all the data values group around a typical or central value.

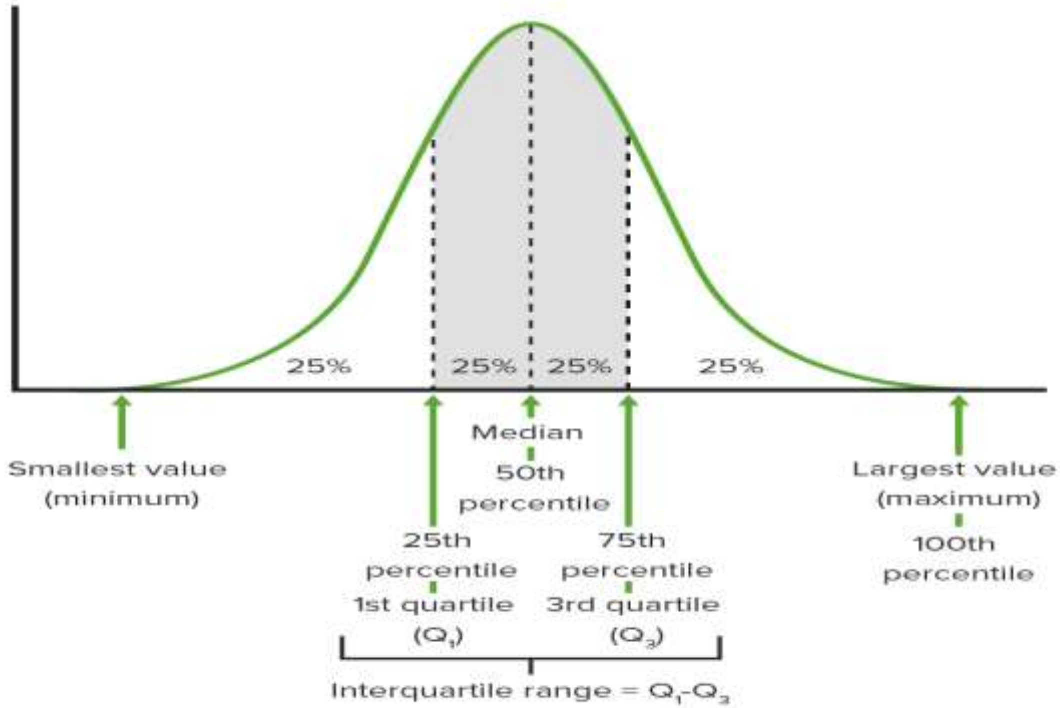
Measures of Shape

- Mean > Median-positive or right Skewness
- Mean = Median- symmetric or zero Skewness
- Mean < Median-Negative of left Skewness



Measures of Dispersion

- Range
- Interquartile range
- Variance
- Standard Deviation
- Coefficient of Variation



Measures of Location

It is a property of the data that they tend to be clustered about a center point.

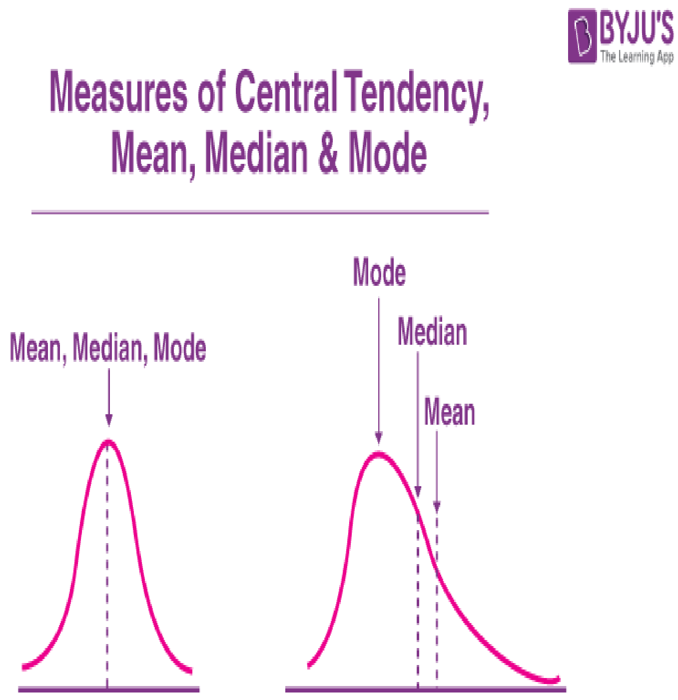
Measures of *central tendency* (i.e., central location) help find the approximate center of the dataset.

Researchers usually do not use the term average, because there are three alternative types of average.

These include the *mean*, the *median*, and the *mode*.

In a perfect world, the mean, median & mode would be the same.

- Mean (generally not part of the data set)
- Median (may be part of the data set)
- Mode (always part of the data set)



Commonly Used Symbols

For a Sample

$\overline{x}$ sample mean

s^2 sample variance

s sample standard deviation

For a Population

μ population mean

σ^2 population variance

σ population standard deviation

Central Tendency Measures		
Measure	**Formula**	**Description**
Mean	∑x/n	Balance Point
Median	n+1/2 Position	Middle Value when ordered
Mode	None	Most frequent

The Mean

- The sample mean is the sum of all the observations (ΣX_i) divided by the number of observations (n):

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \quad \text{where } \Sigma X_i = X_1 + X_2 + X_3 + X_4 + \ldots + X_n$$

- **Example.** 1, 2, 2, 4, 5, 10. Calculate the mean.
  Note: n = 6 (six observations)

$$\Sigma X_i = 1 + 2 + 2 + 4 + 5 + 10 = 24$$

$$\bar{X} = 24 / 6 = 4.0$$

General Formula--Population Mean

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

μ = population mean

Σ = summation sign

x_i = value of element i of the sample

N = population size

Notes on Sample Mean $\overline{X}$

Formula

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

Summation Sign

- In the formula to find the mean, we use the "summation sign" — $\sum$
- This is just mathematical shorthand for "add up all of the observations"

$$\sum_{i=1}^{n} X_i = X_1 + X_2 + X_3 + \ldots\ldots. + X_n$$

Notes on Sample Mean

Also called *sample average* or *arithmetic mean*

Mean for the sample = *X or M, Mean for population = mew (μ)*

Uniqueness: For a given set of data there is one and only one mean.

Simplicity: The mean is easy to calculate.

Sensitive to extreme values

The Mean

- **Example.**

For the data: 1, 1, 1, 1, 51. Calculate the mean.
Note: $n = 5$ (five observations)

$\Sigma X_i = 1 + 1 + 1 + 1 + 51 = 55$
$\bar{X} = 55 / 5 = 11.0$

- Here we see that the mean is affected by extreme values.

The Median

The median is the middle value of the *ordered* data

To get the median, we must first rearrange the data into an ***ordered array*** (in ascending or descending order).

Generally, we order the data from the lowest value to the highest value.

Therefore, the median is the data value such that half of the observations are larger and half are smaller. It is also the 50th percentile.

If n is odd, the median is the middle observation of the ordered array. If n is even, it is midway between the *two* central observations.

The Median

Example:

0	2	3	5	20	99	100

Note: Data has been ordered from lowest to highest. Since n is odd (n=7), the median is the (n+1)/2 ordered observation, or the 4th observation.

Answer: The median is 5.

The mean and the median are unique for a given set of data. There will be exactly one mean and one median.

Unlike the mean, the median is not affected by extreme values.

Q: What happens to the median if we change the 100 to 5,000? Not a thing, the median will still be 5. Five is still the middle value of the data set.

The Median

Example:

10	20	30	40	50	60

Note: Data has been ordered from lowest to highest. Since n is even (n=6), the median is the (n+1)/2 ordered observation, or the 3.5th observation, *i.e.,* the average of observation 3 and observation 4.

Answer: The median is 35.

The Median

- The median has 3 interesting characteristics:
  - 1. The median is not affected by extreme values, only by the number of observations.
  - 2. Any observation selected at random is just as likely to be greater than the median as less than the median.
  - 3. Summation of the absolute value of the differences about the median is a minimum:

$$\sum_{i=0}^{n} |X_i - Median| = minimum$$

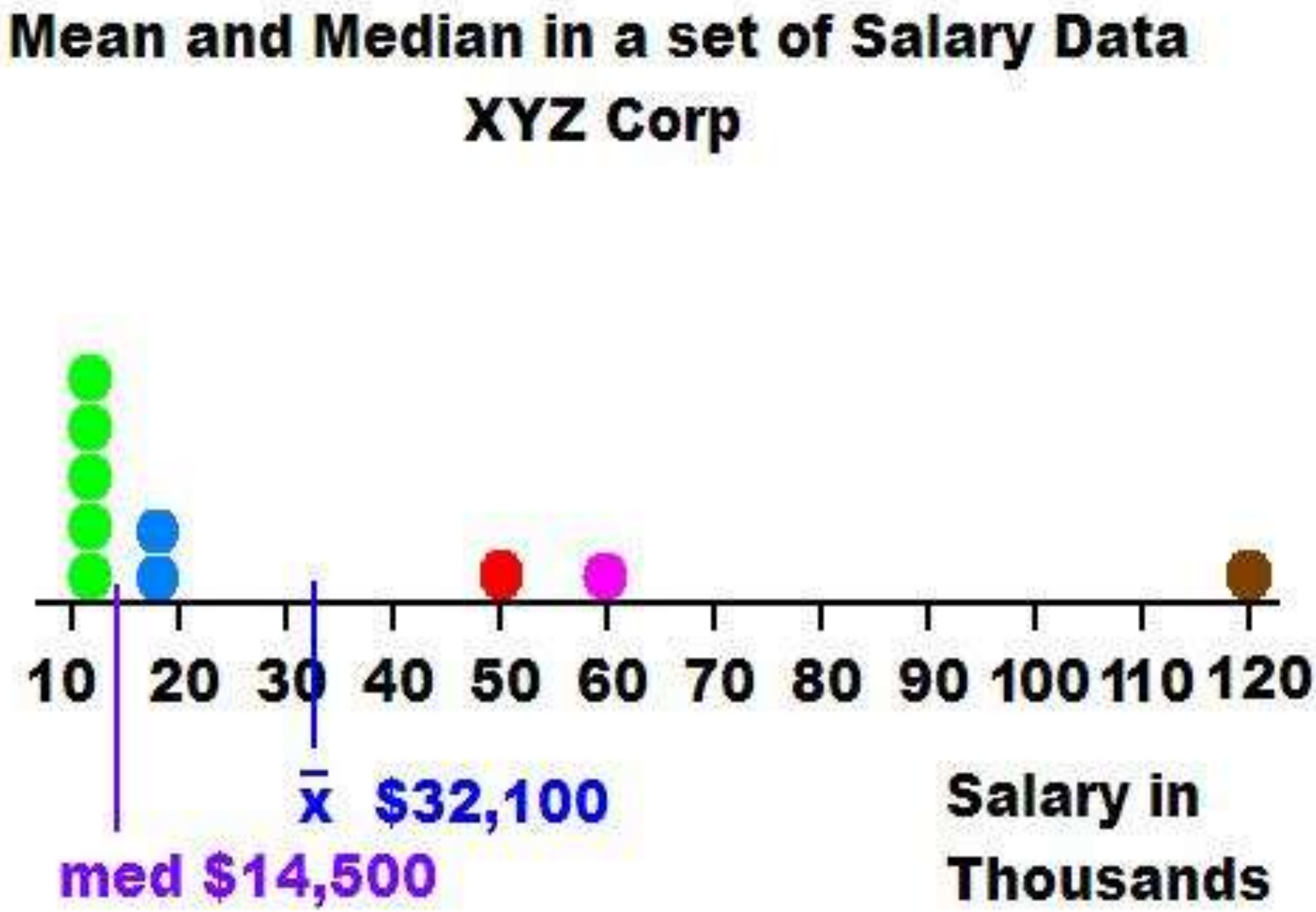
Mean vs. Median

Advantage: The median is less affected by extreme values.

Disadvantage:

- The median takes no account of the precise magnitude of most of the observations and is therefore less efficient than the mean
- If two groups of data are pooled the median of the combined group can not be expressed in terms of the medians of the two original groups but the sample mean can.

$$\bar{x}_{pooled} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$



Mean and Median in a set of Salary Data
XYZ Corp
10 20 30 40 50 60 70 80 90 100 110 120
x̄ $32,100
med $14,500
Salary in Thousands

The Mode

The mode is the value of the data that occurs with the greatest frequency.

Unstable index: values of modes tend to fluctuate from one sample to another drawn from the same population

Example. 1, 1, 1, 2, 3, 4, 5

Answer. The mode is 1 since it occurs three times. The other values each appear only once in the data set.

Example. 5, 5, 5, 6, 8, 10, 10, 10.

Answer. The mode is: 5, 10.

There are two modes. This is a ***bi-modal*** dataset.

The Mode

The mode is different from the mean and the median in that those measures always exist and are always unique. For any numeric data set there will be one mean and one median.

The mode may not exist.

- Data: 1, 2, 3, 4, 5, 6, 7, 8, 9, 0
- Here you have 10 observations and they are all different.

The mode may not be unique.

- Data: 0, 1, 1, 2, 2, 3, 3, 4, 4, 5, 5, 6, 6, 7
- Mode = 1, 2, 3, 4, 5, and 6. There are ***six*** modes.

Comparison of the Mode, the Median, and the Mean

In a normal distribution, the mode , the median, and the mean have the same value.

The mean is the widely reported index of central tendency for variables measured on an interval and ratio scale.

The mean takes each and every score into account.

It also the most stable index of central tendency and thus yields the most reliable estimate of the central tendency of the population.

Comparison of the Mode, the Median, and the Mean

The mean is always pulled in the direction of the long tail, that is, in the direction of the extreme scores.

For the variables that positively skewed (like income), the mean is higher than the mode or the median. For negatively skewed variables (like age at death) the mean is lower.

When there are extreme values in the distribution (even if it is approximately normal), researchers sometimes report means that have been adjusted for outliers.

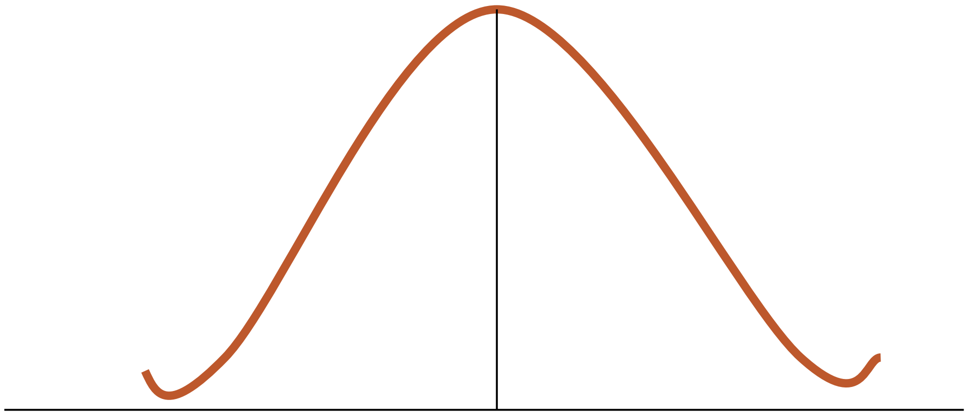
To adjust means one must discard a fixed percentage (5%) of the extreme values from either end of the distribution.

Distribution Characteristics

Mode: Peak(s)

Median: Equal areas point

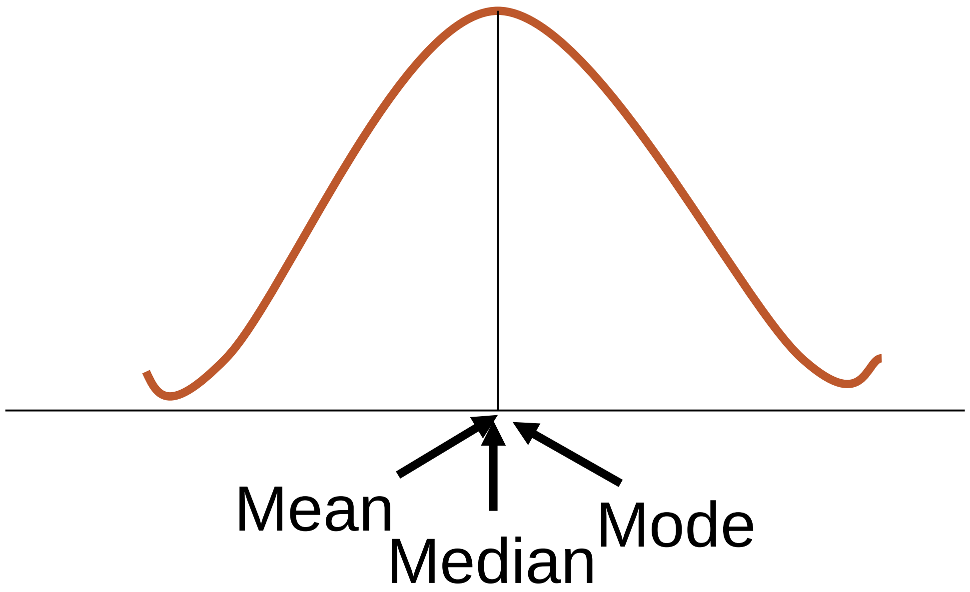
Mean: Balancing point



Shapes of Distributions

Symmetric (Right and left sides are mirror images)

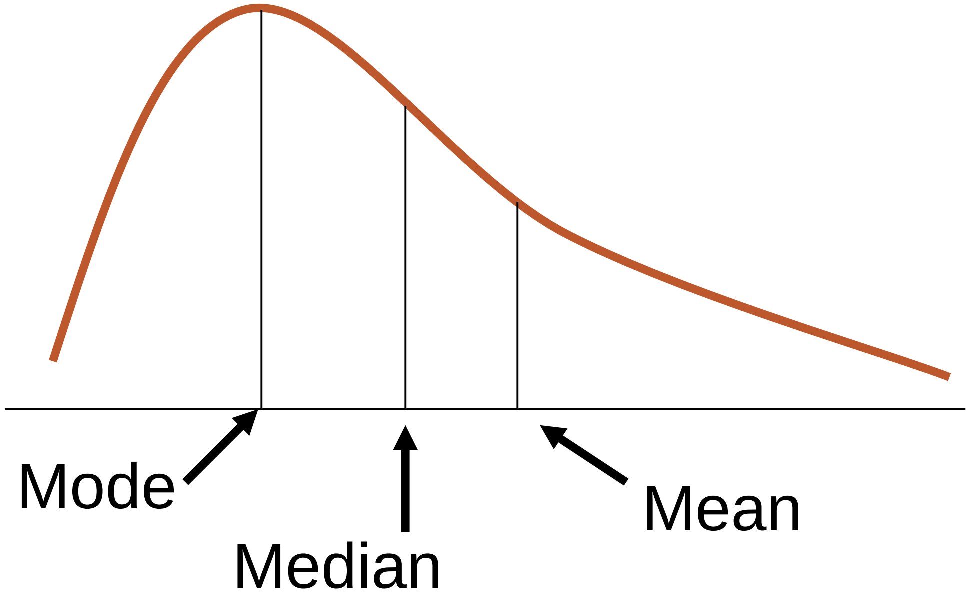
- Left tail looks like right tail
- Mean = Median = Mode



Shapes of Distributions

Right skewed (positively skewed)

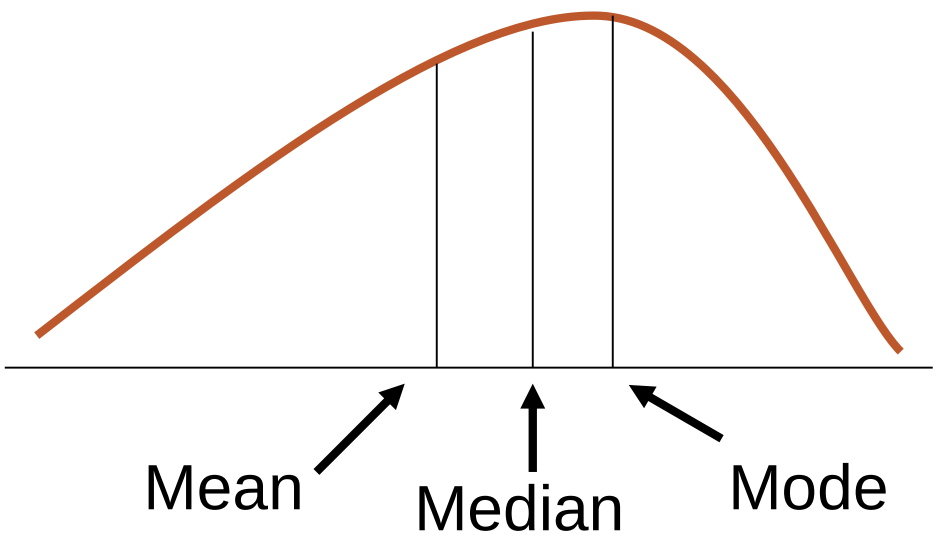
- Long right tail
- Mean > Median



Shapes of Distributions

Left skewed (negatively skewed)

- Long left tail
- Mean < Median



Quantiles

Measures of non-central location used to summarize a set of data

Examples of commonly used quantiles:

- Quartiles
- Quintiles
- Deciles
- Percentiles

Quartiles

Quartiles split a set of ordered data into four parts.

- Imagine cutting a chocolate bar into four equal pieces… How many cuts would you make? (yes, 3!)

Q_1 is the First Quartile

- 25% of the observations are smaller than Q_1 and 75% of the observations are larger

Q_2 is the Second Quartile

- 50% of the observations are smaller than Q_2 and 50% of the observations are larger. Same as the Median. It is also the 50th percentile.

Q_3 is the Third Quartile

- 75% of the observations are smaller than $Q_3$and 25% of the observations are larger

Quartiles

A quartile, like the median, either takes the value of one of the observations, or the value halfway between two observations.

- If n/4 is an integer, the first quartile (Q1) has the value halfway between the (n/4)th observation and the next observation.
- If n/4 is not an integer, the first quartile has the value of the observation whose position corresponds to the next highest integer.

210
220
225
225
225
235
240
250
270
280

$Q_1 = 225$, the 3rd observation

Q_2 = Median - 230

Q_3=250, 3rd observation from the bottom

- The method we are using is an approximation. If you solve this in MS Excel, which relies on a formula, you may get an answer that is slightly different.

Exercise

- Computer Sales (n = 12 salespeople)
  Original Data: 3, 10, 2, 5, 9, 8, 7, 12, 10, 0, 4, 6
  Compute the mean, median, mode, quartiles.

  First order the data:
  0, 2, 3, 4, 5, 6, 7, 8, 9, 10, 10, 12

  $\Sigma X_i = 76$

  $\overline{X} = 76 / 12 = 6.33$ computers sold

  Median = 6.5 computers

  Mode = 10 computers

  $Q_1 = 3.5$ computers, $Q_3 = 9.5$ computers

Other Quantiles

Similar to what we just learned about quartiles, where 3 quartiles split the data into 4 equal parts,

- There are 9 *deciles* dividing the distribution into 10 equal portions (tenths).
- There are four *quintiles* dividing the population into 5 equal portions.
- … and 99 percentiles (next slide)

In all these cases, the convention is the same. The point, be it a quartile, decile, or *percentile*, takes the value of one of the observations or it has a value halfway between two adjacent observations. It is never necessary to split the difference between two observations more finely.

Percentiles

We use 99 *percentiles* to divide a data set into 100 equal portions.

Percentiles are used in analyzing the results of standardized exams. For instance, a score of 40 on a standardized test might seem like a terrible grade, but if it is the 99th percentile, don't worry about telling your parents. ☺

Which percentile is Q1? Q2 (the median)? Q3?

We will always use computer software to obtain the percentiles.

Some Exercises

Data (n=16):

1, 1, 2, 2, 2, 2, 3, 3, 4, 4, 5, 5, 6, 7, 8, 10

Compute the mean, median, mode, quartiles.

Answer.

1 1 2 2 ¦ 2 2 3 3 ¦ 4 4 5 5 ¦ 6 7 8 10

Mean = 65/16 = 4.06

Median = 3.5

Mode = 2

Q_1 = 2

Q_2 = Median = 3.5

Q_3 = 5.5

Exercise: # absences

Data – number of absences (n=13) :

0, 5, 3, 2, 1, 2, 4, 3, 1, 0, 0, 6, 12

Compute the mean, median, mode, quartiles.

Answer. First order the data:

0, 0, 0, ¦ 1, 1, 2, 2, 3, 3, 4, ¦ 5, 6, 12

Mean = 39/13 = 3.0 absences

Median = 2 absences

Mode = 0 absences

Q_1 = .5 absences

Q_3 = 4.5 absences

Exercise: Reading level

Data: Reading Levels of 16 eighth graders.
5, 6, 6, 6, 5, 8, 7, 7, 7, 8, 10, 9, 9, 9, 9, 9

Answer. First, order the data:

5 5 6 6 ┆ 6 7 7 7 ┆ 8 8 9 9 ┆ 9 9 9 10

Sum=120.

Mean= 120/16 = 7.5 This is the average reading level of the 16 students.

Median = Q_2 = 7.5

Q_1 = 6, Q_3 = 9

Mode = 9

Measures of Dispersion

It refers to how spread out the scores are.

In other words, how similar or different participants are from one another on the variable. It is either homogeneous or heterogeneous sample.

Why do we need to look at measures of dispersion?

Consider this example:

A company is about to buy computer chips that must have an average life of 10 years. The company has a choice of two suppliers. Whose chips should they buy? They take a sample of 10 chips from each of the suppliers and test them. See the data on the next slide.

Measures of Dispersion

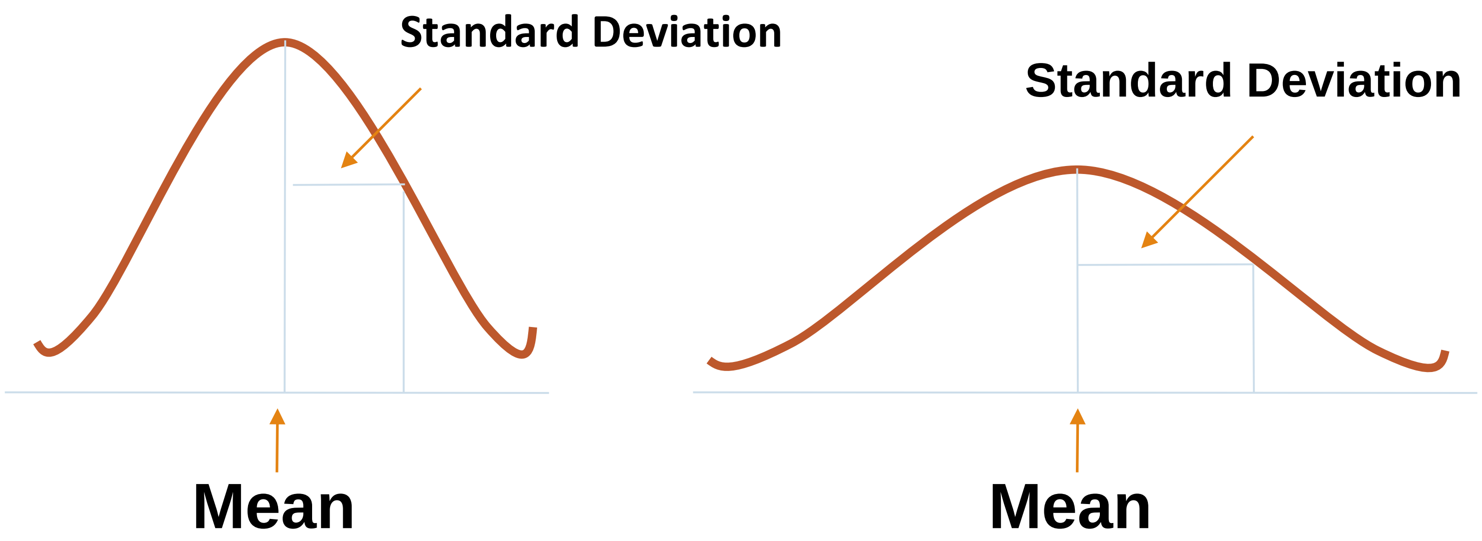
We see that supplier B's chips have a longer average life.

However, what if the company offers a 3-year warranty?

Then, computers manufactured using the chips from supplier A will have no returns while using supplier B will result in 4/10 or 40% returns.

Supplier A chips (life in years)	Supplier B chips (life in years)
11	170
11	1
10	1
10	160
11	2
11	150
11	150
11	170
10	2
12	140
$\overline{X}_A = 10.8$ years	$\overline{X}_B = 94.6$ years
$\text{Median}_A = 11$ years	$\text{Median}_B = 145$ years
$s_A = 0.63$ years	$s_B = 80.6$ years
$\text{Range}_A = 2$ years	$\text{Range}_B = 169$ years

Normal Distribution



Measures of Dispersion

We will study these five measures of dispersion

- Range
- Interquartile Range
- Standard Deviation
- Variance
- Coefficient of Variation
- Relative Standing.

The Range

Is the simplest measure of variability, is the difference between the highest score and the lowest score in the distribution.

In research, the range is often shown as the minimum and maximum value, without the abstracted difference score.

It provides a quick summary of a distribution's variability.

It also provides useful information about a distribution when there are extreme values.

The range has two values, it is highly unstable.

The Range

Range = Largest Value – Smallest Value

Example: 1, 2, 3, 4, 5, 8, 9, 21, 25, 30

Answer: Range = 30 – 1 = 29.

Pros:

- Easy to calculate

Cons:

- Value of range is only determined by two values
- The interpretation of the range is difficult.
- One problem with the range is that it is influenced by extreme values at either end.

Standard Deviation

- The standard deviation, *s*, measures a kind of "average" deviation about the mean. It is not really the "average" deviation, even though we may think of it that way.

- Why can't we simply compute the average deviation about the mean, if that's what we want?

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

- If you take a simple mean, and then add up the deviations about the mean, as above, this sum will be equal to 0. Therefore, a measure of "average deviation" will not work.

Standard Deviation

- Instead, we use:

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

- This is the "definitional formula" for standard deviation.
- The standard deviation has lots of nice properties, including:
  - By squaring the deviation, we eliminate the problem of the deviations summing to zero.
  - In addition, this sum is a minimum. No other value subtracted from X and squared will result in a smaller sum of the deviation squared. This is called the "least squares property."
- Note we divide by (n-1), not n. This will be referred to as a loss of one degree of freedom.

Standard Deviation

The smaller the standard deviation, the better is the mean as the summary of a typical score. E.g. 10 people weighted 150 pounds, the SD would be zero, and the mean of 150 would communicate perfectly accurate information about all the participants wt. Another example would be a heterogeneous sample 5 people 100 pounds and another five people 200 pounds. The mean still 150, but the SD would be 52.7.

In normal distribution there are 3 SDs above the mean and 3 SDs below the mean.

Standard Deviation

- **Example.** Two data sets, X and Y. Which of the two data sets has greater variability? Calculate the standard deviation for each.

We note that both sets of data have the same mean:

$\overline{X} = 3$

$\overline{Y} = 3$

X_i	Y_i
1	0
2	0
3	0
4	5
5	10

(continued...)

Standard Deviation

$S_X = \sqrt{\frac{10}{4}} = 1.58$

X	$\overline{X}$	$(X-\overline{X})$	$(X-\overline{X})^2$
1	3	-2	4
2	3	-1	1
3	3	0	0
4	3	1	1
5	3	2	4
		$\sum=0$	10

$S_Y = \sqrt{\frac{80}{4}} = = 4.47$

Y	$\overline{Y}$	$(Y-\overline{Y})$	$(Y-\overline{Y})^2$
0	3	-3	9
0	3	-3	9
0	3	-3	9
5	3	2	4
10	3	7	49
		$\sum=0$	80

[Check these results with your calculator.]

Standard Deviation: N vs. (n-1)

- Note that $\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}}$ and $s = \sqrt{\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{n-1}}$

- You divide by N only when you have taken a census and therefore know the population mean. This is rarely the case.

- Normally, we work with a sample and calculate sample measures, like the sample mean and the sample standard deviation:

- The reason we divide by n−1 instead of n is to assure that *s* is an *unbiased* estimator of σ.
  - We have taken a shortcut: in the second formula we are using the sample mean, $\overline{X}$, a statistic, in lieu of μ, a population parameter. Without a correction, this formula would have a tendency to understate the true standard deviation. We divide by n−1, which increases *s*. This makes it an *unbiased estimator* of σ.
  - We will refer to this as "losing one degree of freedom" (to be explained more fully later on in the course).

Variance

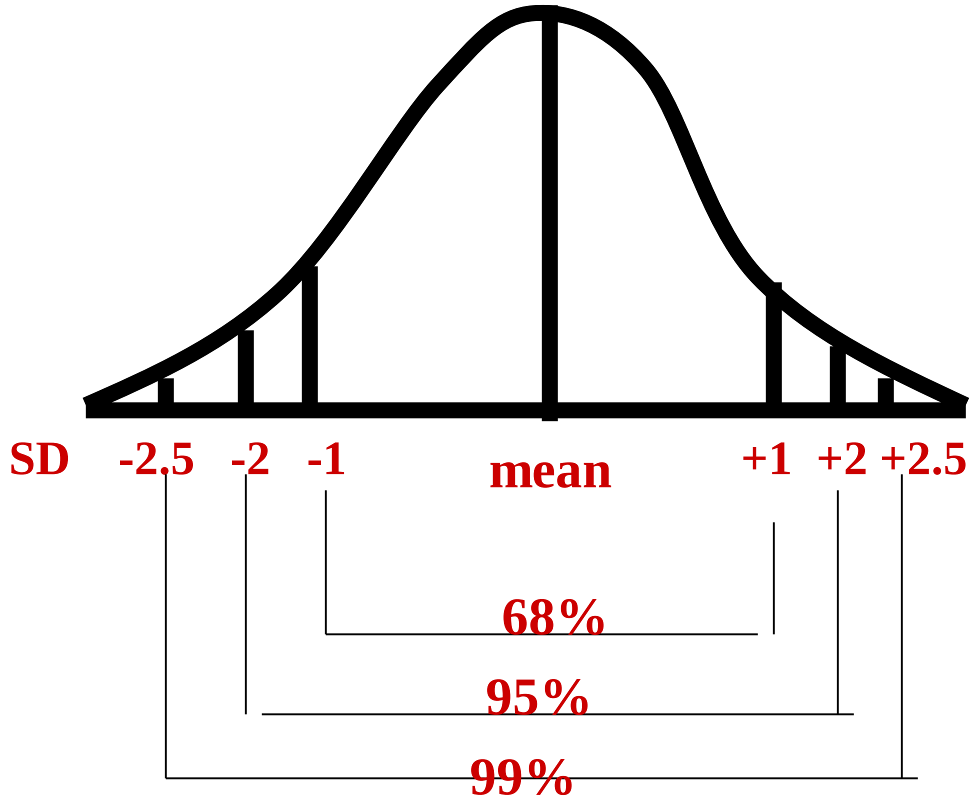
- The variance, s^2, is the standard deviation (s) squared. Conversely, $s = \sqrt{variance}$.

Definitional formula: $s^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$

Computational formula: $s^2 = \frac{\sum_{i=1}^{n}(X_i)^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}}{n-1}$

This is what computer software (e.g., MS Excel or your calculator key) uses.

Relationship between SD and frequency distribution



Coefficient of Variation (CV)

The problem with s^2 and s is that they are both, like the mean, in the "original" units.

This makes it difficult to compare the variability of two data sets that are in different units or where the magnitude of the numbers is very different in the two sets. For example,

- Suppose you wish to compare two stocks and one is in dollars and the other is in yen; if you want to know which one is more volatile, you should use the coefficient of variation.
- It is also not appropriate to compare two stocks of vastly different prices even if both are in the same units.
- The standard deviation for a stock that sells for around \$300 is going to be very different from one with a price of around \$0.25.

The *coefficient of variation* will be a better measure of dispersion in these cases than the standard deviation (see example on the next slide).

$$CV = \frac{s}{\bar{\bar{X}}}(100\%)$$

Coefficient of Variation (CV)

- $CV = \frac{s}{\overline{\overline{X}}}(100\%)$

CV is in terms of a percent. What we are in effect calculating is what percent of the sample mean is the standard deviation. If CV is 100%, this indicates that the sample mean is equal to the sample standard deviation. This would demonstrate that there is a great deal of variability in the data set. 200% would obviously be even worse.

Example: Stock Prices

Which stock is more volatile?

Closing prices over the last 8 months:

$$CV_A = \frac{\$1.62}{\$1.70} \times 100\% = 95.3\%$$

$$CV_B = \frac{\$11.33}{\$188.88} \times 100\% = 6.0\%$$

	Stock A	Stock B
JAN	\$1.00	\$180
FEB	1.50	175
MAR	1.90	182
APR	.60	186
MAY	3.00	188
JUN	.40	190
JUL	5.00	200
AUG	.20	210
Mean	\$1.70	\$188.88
s^2	2.61	128.41
s	\$1.62	\$11.33

Answer: The standard deviation of B is higher than for A, but A is more volatile:

Exercise: Test Scores

- Data (n=10): 0, 0, 40, 50, 50, 60, 70, 90, 100, 100
  Compute the mean, median, mode, quartiles (Q1, Q2, Q3), range, interquartile range, variance, standard deviation, and coefficient of variation. We shall refer to all these as the descriptive (or summary) statistics for a set of data.

Answer. First order the data:

0, 0, 40, 50, 50 ¦ 60, 70, 90, 100, 100

- Mean: $\Sigma X_i = 560$ and $n = 10$, so $\overline{X} = 560/10 = 56$.
  Median = Q_2 = 55
- Q_1 = 40 ; Q_3 = 90 (Note: Excel gives these as Q_1 = 42.5, Q_3 = 85.)
- Mode = 0, 50, 100
  Range = 100 – 0 = 100
- IQR = 90 – 40 = 50
- s^2 = 11,840/9 = 1315.5
- $s = \sqrt{1315.5} = 36.27$
- CV = (36.27/56) x 100% = 64.8%

The Interquartile range (IQR)

The Interquartile range (IQR) is the score at the 75th percentile or 3rd quartile (Q3) minus the score at the 25th percentile or first quartile (Q1). Are the most used to define outliers.

It is not sensitive to extreme values.

Inter-Quartile Range (IQR)

$IQR = Q_3 - Q_1$

Example (n = 15):

0, 0, 2, 3, 4, 7, 9, 12, 17, 18, 20, 22, 45, 56, 98

$Q_1 = 3$, $Q_3 = 22$

IQR = 22 – 3 = 19 (Range = 98)

This is basically the range of the central 50% of the observations in the distribution.

Problem: The Interquartile range does not take into account the variability of the *total* data (only the central 50%). We are "throwing out" half of the data.

Five Number Summary

- When examining a distribution for shape, sometime the five number summary is useful:

  Smallest | Q1 | Median | Q3 | Largest

- Example:

2	3	8	8	9	10	10	12	15	18	22	63

$\overline{X} = 15$

Smallest Q_1 Median Q_3 Largest

5-number summary: 2 | 8 | 10 | 16.5 | 63

This data is right-skewed.

In right-skewed distributions, the distance from Q_3 to $X_{largest}$ (16.5 to 63) is significantly greater than the distance from $X_{smallest}$ to Q_1(2 to 8).

Standard Scores

There are scores that are expressed in terms of their relative distance from the mean. It provides information not only about rank but also distance between scores.

It often called Z-score.

Z Score

Is a standard score that indicates how many SDs from the mean a particular values lies.

Z = Score of value – mean of scores divided by standard deviation.

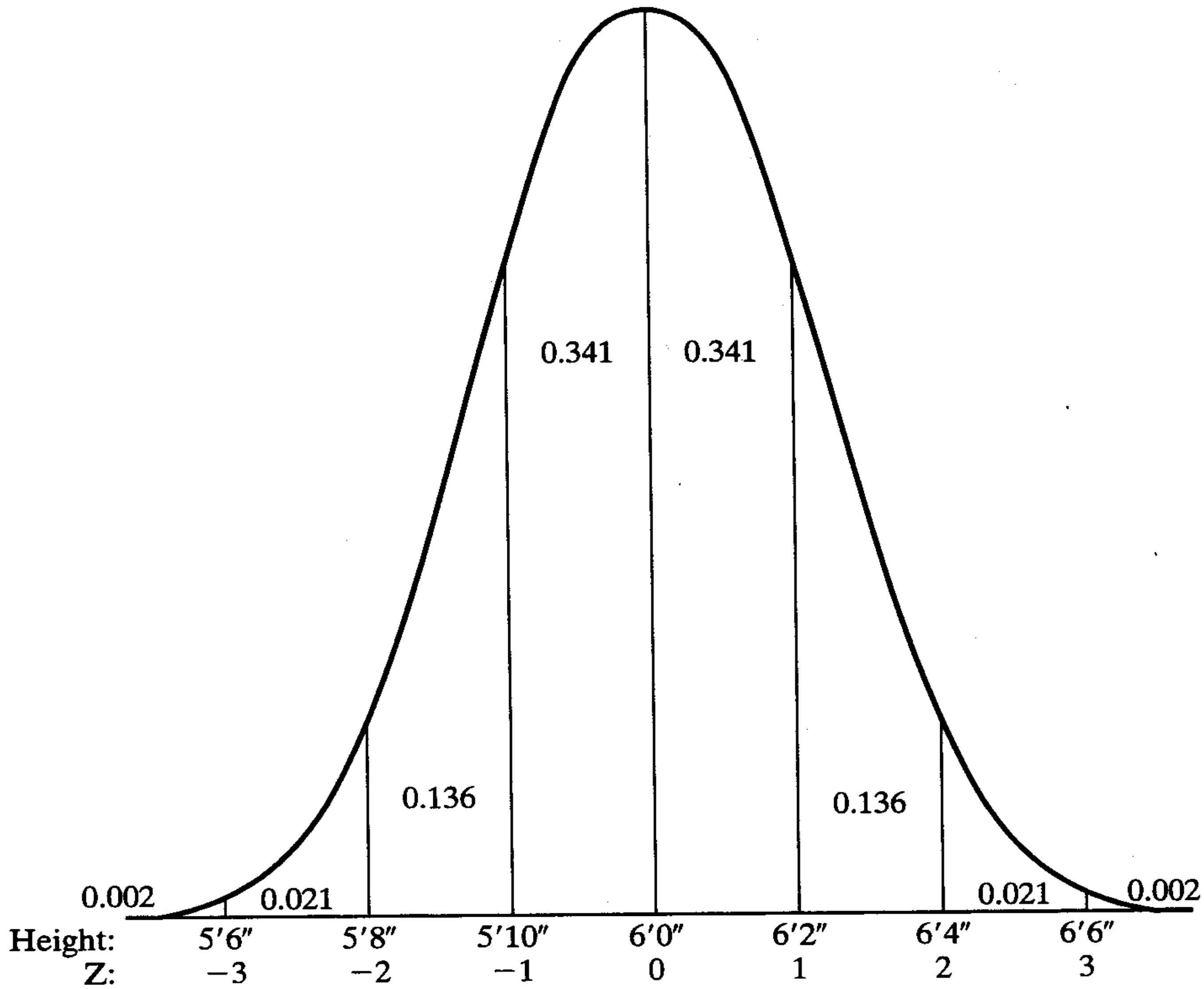
Standard Normal Scores

How many standard deviations away from the mean are you?

Standard Score (Z) =

Observation – mean Standard deviation

"Z" is normal with mean 0 and standard deviation of 1.


0.341
0.341
0.136
0.136
0.002
0.021
0.021
0.002
Height:
5′6″
5′8″
5′10″
6′0″
6′2″
6′4″
6′6″
Z:
−3
−2
−1
0
1
2
3

Standard Normal Scores

A standard score of:

Z = 1: The observation lies one SD above the mean

Z = 2: The observation is two SD above the mean

Z = -1: The observation lies 1 SD below the mean

Z = -2: The observation lies 2 SD below the mean

$$Z = \frac{X - \mu}{\sigma}$$

Standard Normal Scores

Example: Male Blood Pressure,
mean = 125, s = 14 mmHg

- BP = 167 mmHg

$$Z = \frac{167 - 125}{14} = 3.0$$

- BP = 97 mmHg

$$Z = \frac{97 - 125}{14} = -2.0$$

What is the Usefulness of a Standard Normal Score?

It tells you how many SDs (s) an observation is from the mean

Thus, it is a way of quickly assessing how "unusual" an observation is

Example: Suppose the mean *BP* is 125 mmHg, and standard deviation = 14 mmHg

- Is 167 mmHg an unusually high measure?
- If we know Z = 3.0, does that help us?

Standardizing Data: Z-Scores

- We can convert the original scores to new scores with $\overline{X} = 0$ and $s = 1$.
- This will give us a pure number with no units of measurement.
- Any score below the mean will now be negative.
- Any score at the mean will be 0.
- Any score above the mean will be positive.

Standardizing Data: Z-Scores

- To compute the Z–scores:

$$Z = \frac{X - \bar{X}}{s}$$

Example.

Data: 0, 2, 4, 6, 8, 10

$\bar{X} = 30/6 = 5$; $s = 3.74$

X	→	Z
0	$\frac{0-5}{3.74}$	-1.34
2	$\frac{2-5}{3.74}$	-.80
4	$\frac{4-5}{3.74}$	-.27
6	$\frac{6-5}{3.74}$	.27
8	$\frac{8-5}{3.74}$	.80
10	$\frac{10-5}{3.74}$	1.34

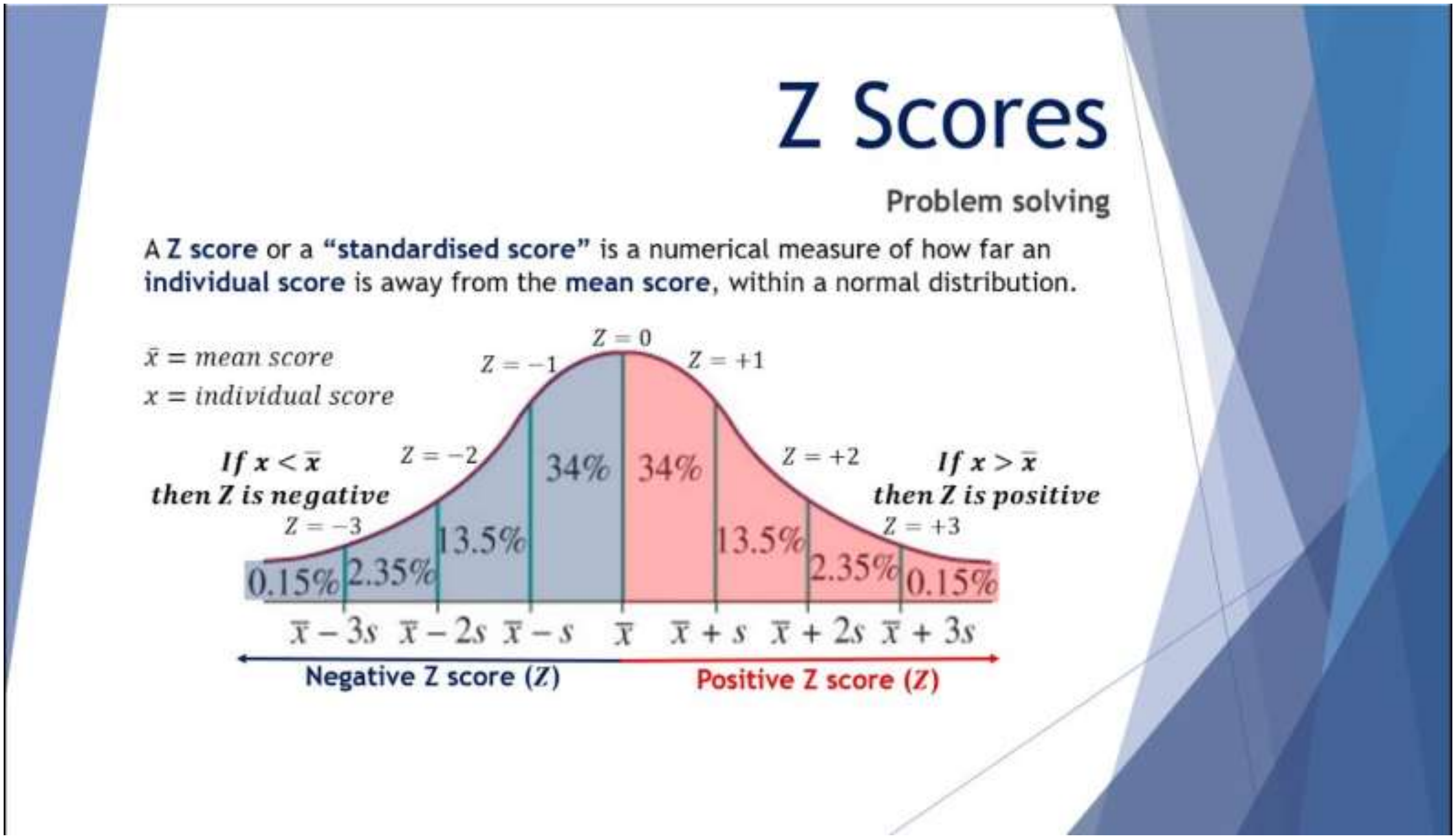
Standardizing Data: Z-Scores

No matter what you are measuring, a Z-score of more than +5 or less than – 5 would indicate a very, very unusual score.

For standardized data, if it is normally distributed, 95% of the data will be between ±2 standard deviations about the mean.

If the data follows a normal distribution,

- 95% of the data will be between -1.96 and +1.96.
- 99.7% of the data will fall between -3 and +3.
- 99.99% of the data will fall between -4 and +4.


Z Scores
Problem solving
A Z score or a "standardised score" is a numerical measure of how far an individual score is away from the mean score, within a normal distribution.
$\bar{x}$ = mean score
x = individual score
If $x < \bar{x}$ then Z is negative
If $x > \bar{x}$ then Z is positive
Z = 0
Z = −1
Z = +1
Z = −2
Z = +2
Z = −3
Z = +3
0.15%
2.35%
13.5%
34%
34%
13.5%
2.35%
0.15%
$\bar{x} - 3s$
$\bar{x} - 2s$
$\bar{x} - s$
$\bar{x}$
$\bar{x} + s$
$\bar{x} + 2s$
$\bar{x} + 3s$
Negative Z score (Z)
Positive Z score (Z)

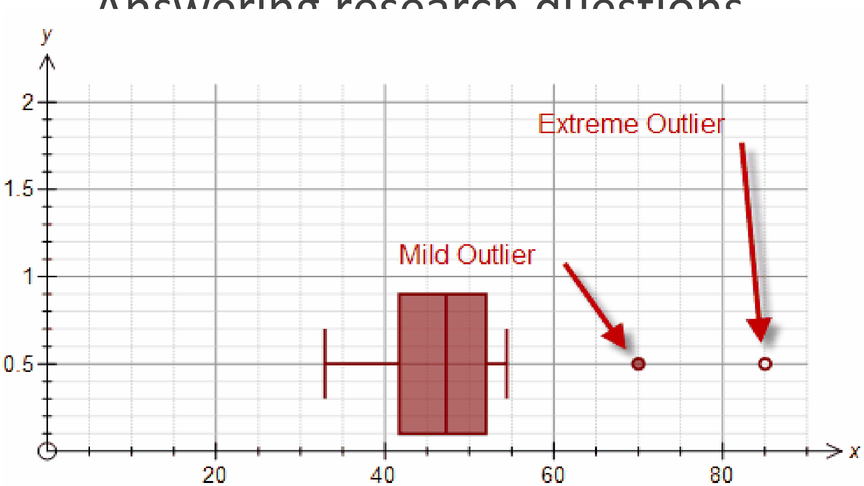
Major reasons for using Index (descriptive statistics)

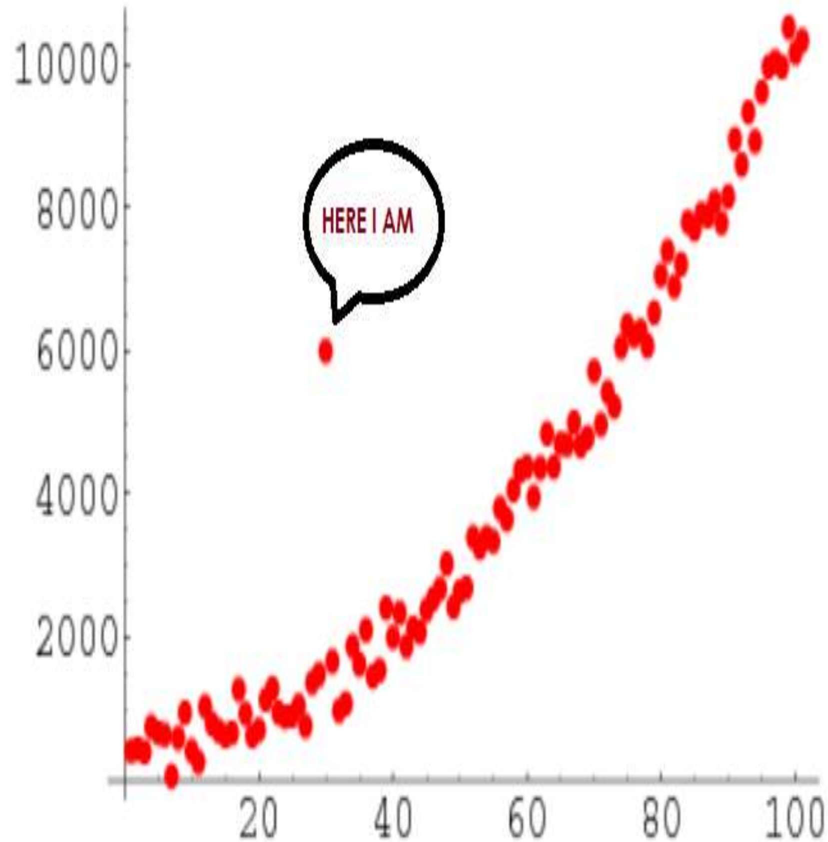
Understanding the data.

Evaluating outliers. Outliers are often identified in relation to the value of a distribution's IQR.

Describe the research sample.

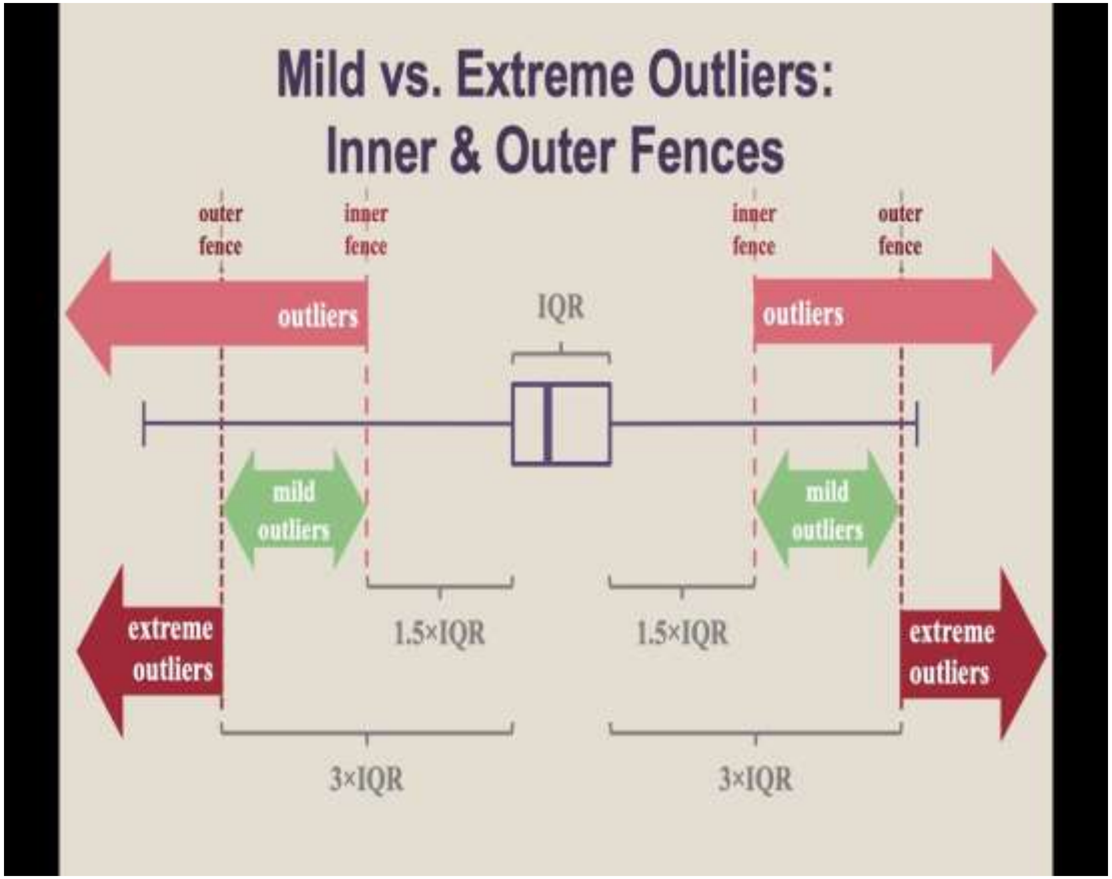
Answering research questions





A mild outlier is a data value that lies between 1.5 and 3 times the IQR below Q1 or above Q3.

Extreme outlier is a data value that is more that three times the IQR below Q1 or above Q3.




THE END OF UNIT 2