

Unit Two

Descriptive Biostatistics

ASSOCIATE PROFESSOR DIANA ARABIAT

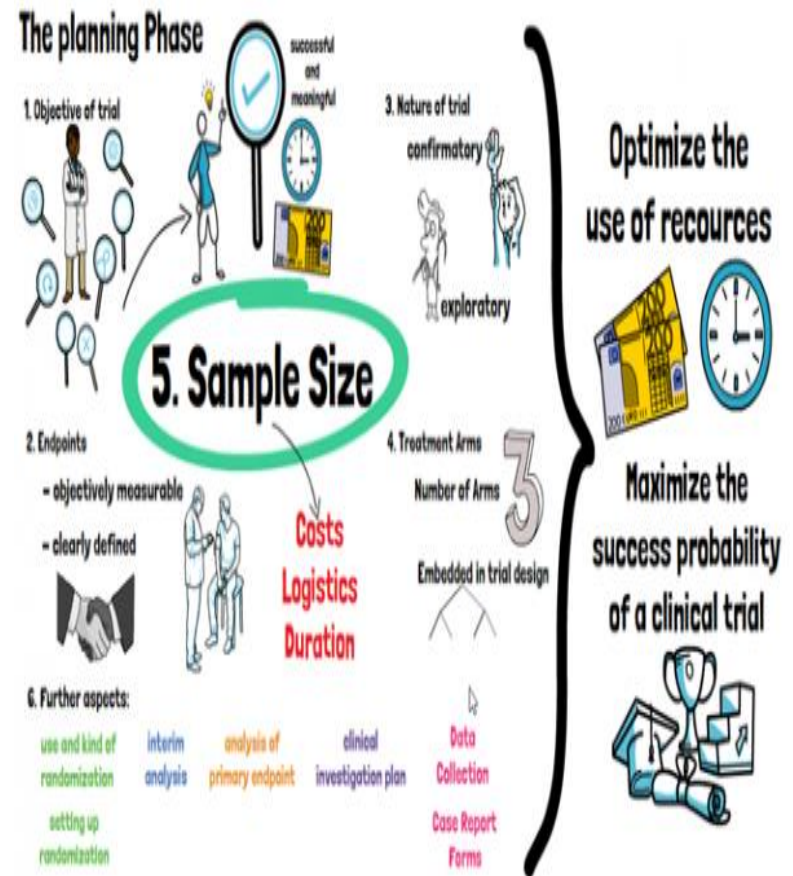


Recap

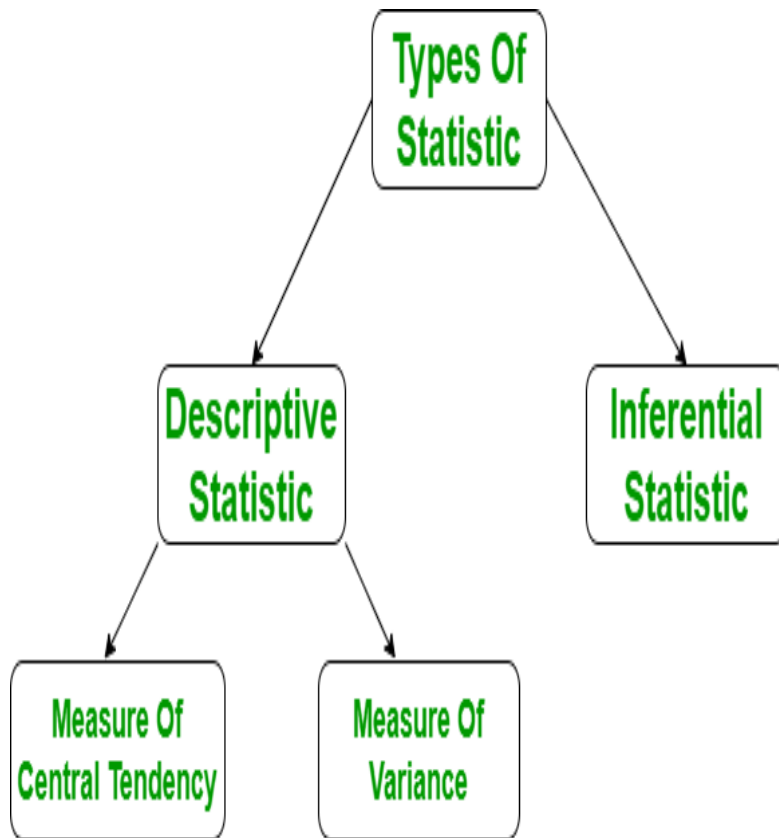
BIOSTATISTICS

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.

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 smartphones 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.

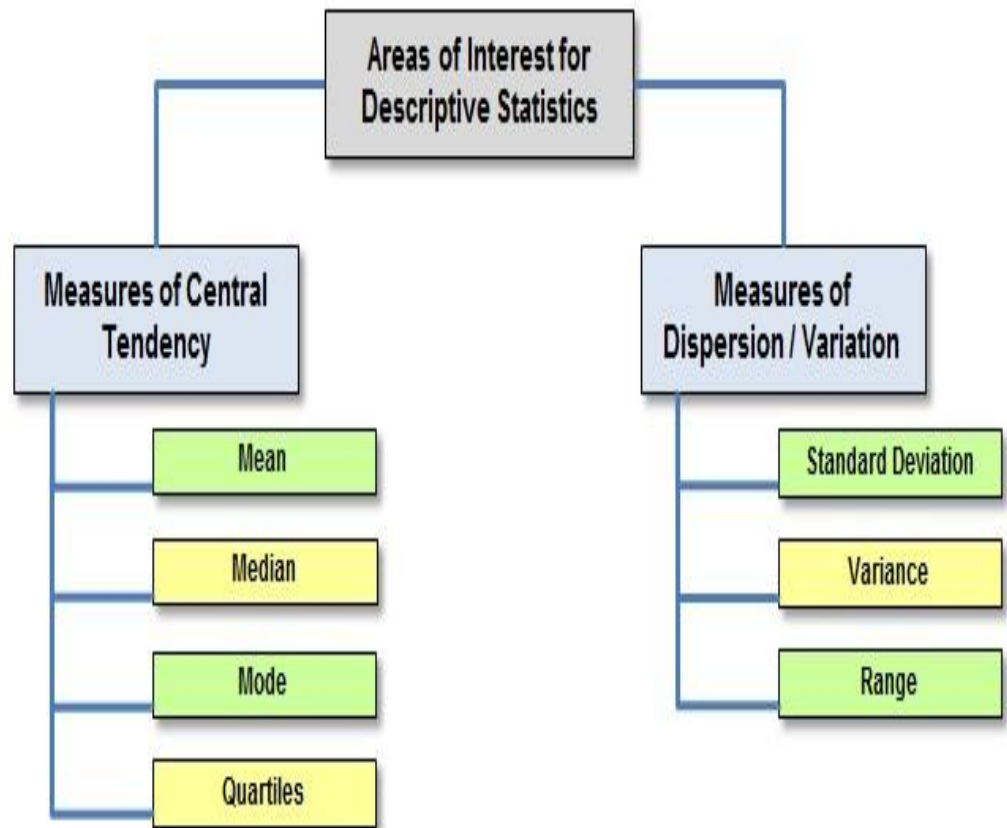
Un-Grouped Data: Data in its original form is called Un-Grouped Data.

Note: Raw data is also called ungrouped data.

Descriptive Measures

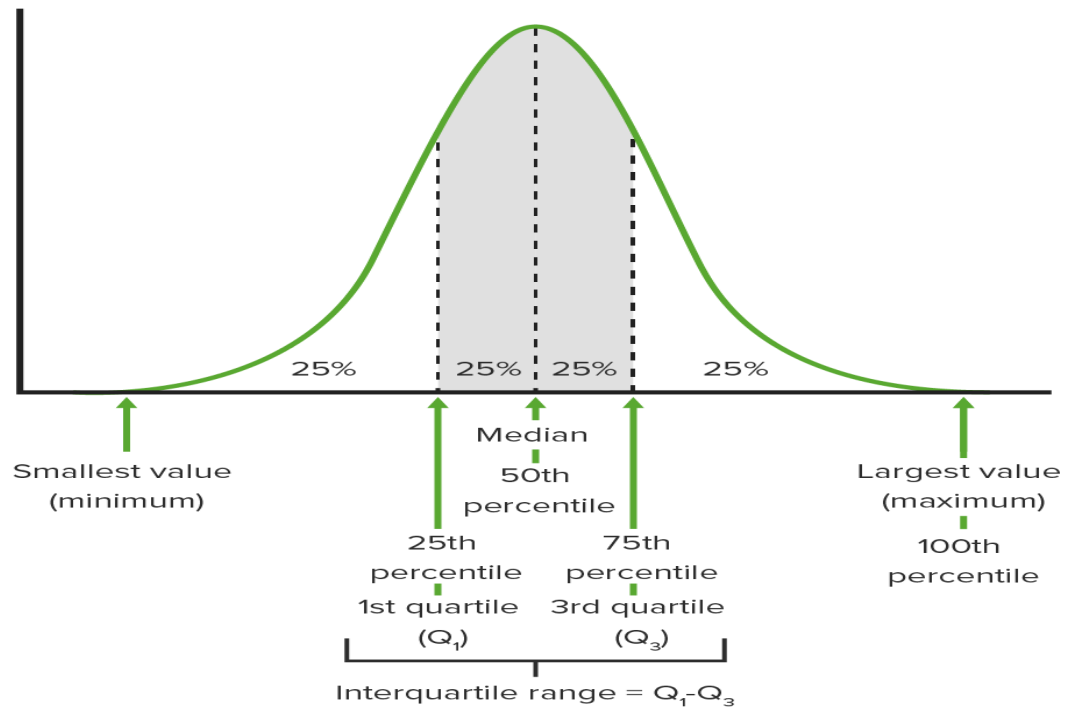
A *descriptive measure* is a single number that is used to describe a set of data.

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



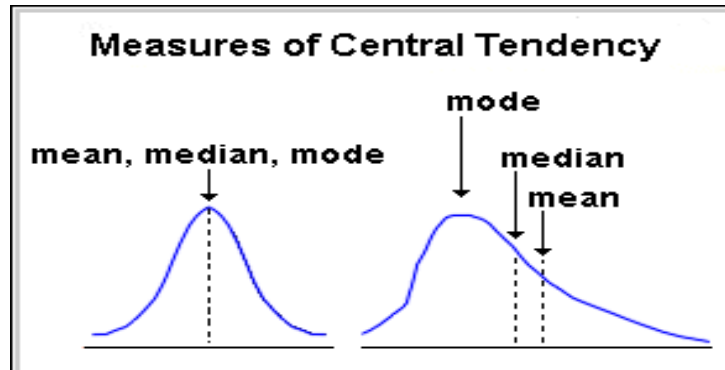
Measures of Dispersion

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



Measures of Location

- Measures of central tendency:
Mean; Median; Mode



- Measures of noncentral tendency
- 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 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.

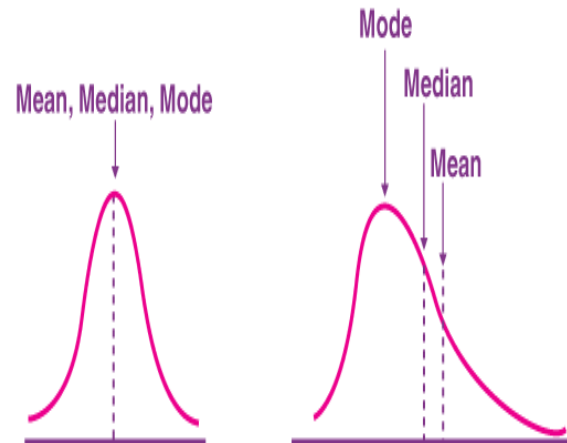
Researchers usually do not use the term average, because there are three alternative types of average.

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)

Measures of Central Tendency, Mean, Median & Mode



Commonly Used Symbols

For a Sample

\bar{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	$\sum x/n$	Balance Point
Median	$n+1/2$ Position	Middle Value when ordered
Mode	None	Most frequent

Notes on Sample Mean \bar{X}

Formula

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

Summation Sign

- In the formula to find the mean, we use the “summation sign” — Σ
- This is just mathematical shorthand for “add up all of the observations”

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

Notes on Sample Mean

Also called *sample average* or *arithmetic mean*

Mean for the sample = \bar{X} or \bar{M} , Mean for population = 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 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

- ▶ 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:

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

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.

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

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.

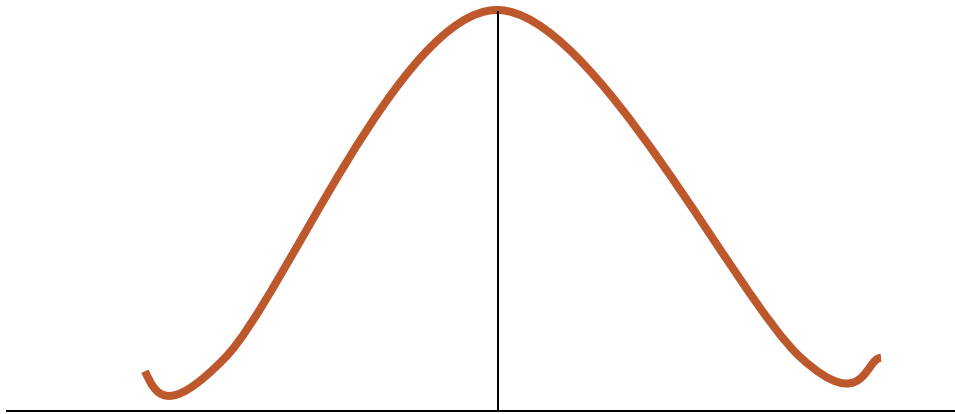
To adjust means one must discard a fixed percentage (5%) of the extreme values from either end of the distribution.

Distribution Characteristics

Mode: Peak(s)

Median: Equal areas point

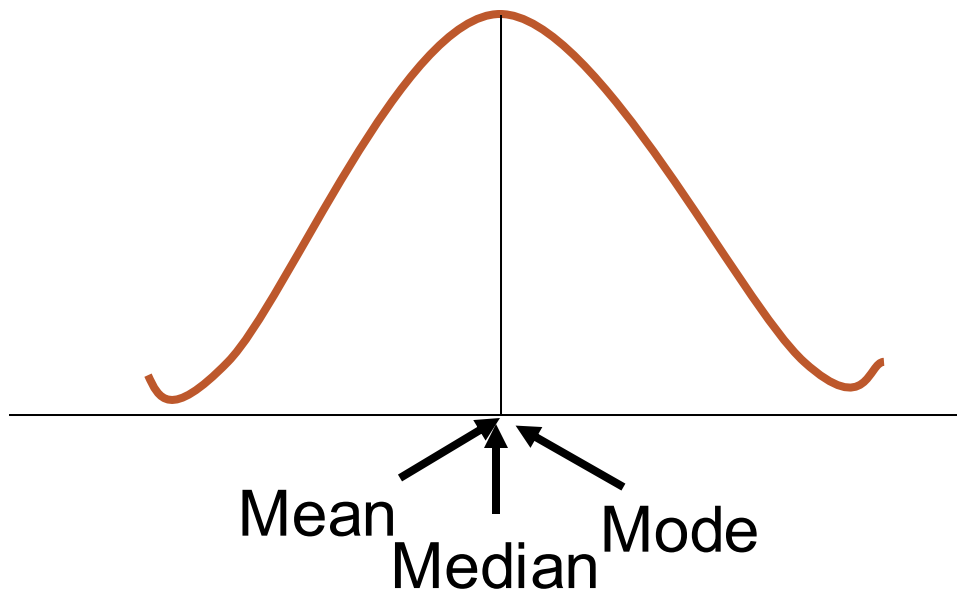
Mean: Balancing point



Shapes of Distributions

Symmetric (Right and left sides are mirror images)

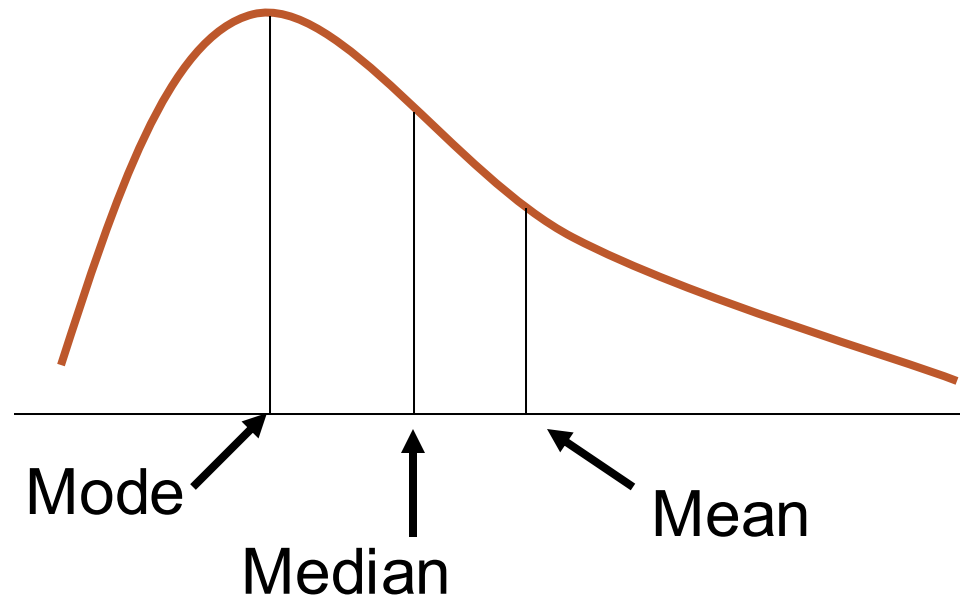
- Left tail looks like right tail
- Mean = Median = Mode



Shapes of Distributions

Right skewed (positively skewed)

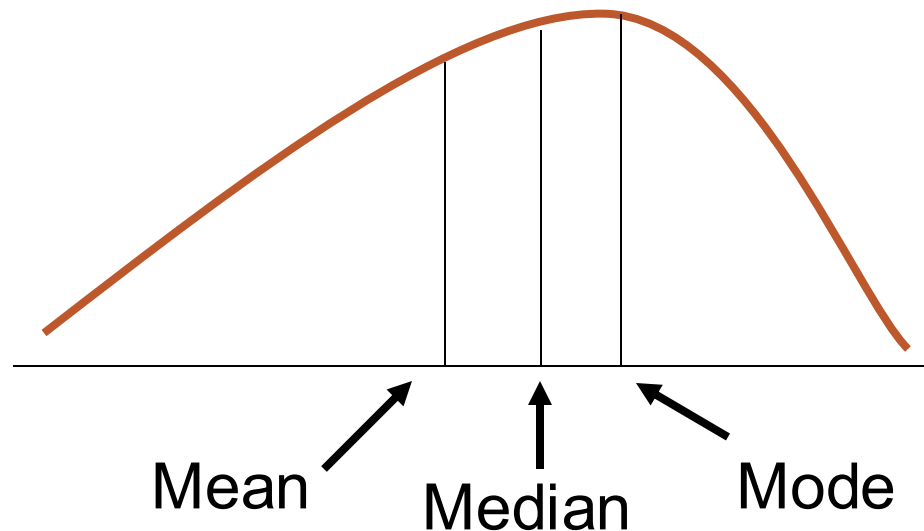
- Long right tail
- $\text{Mean} > \text{Median}$



Shapes of Distributions

Left skewed (negatively skewed)

- Long left tail
- $\text{Mean} < \text{Median}$



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 (Q_1) 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

$Q_2 = \text{Median} = 230$

$Q_3 = 250$, 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.

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.

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.

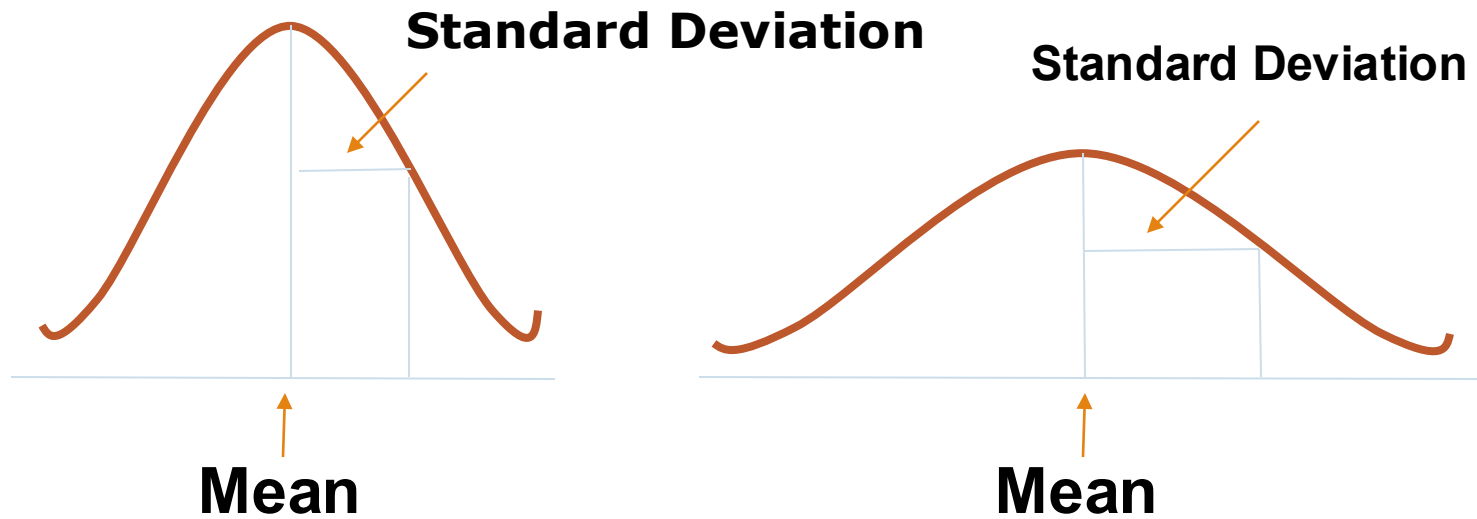
Supplier A chips (life in years)	Supplier B chips (life in years)
11	170
11	1
10	1
10	160
11	2
11	150
11	150
11	170
10	2
12	140
$\bar{X}_A = 10.8$ years	$\bar{X}_B = 94.6$ years
Median _A = 11 years	Median _B = 145 years
$s_A = 0.63$ years	$s_B = 80.6$ years
Range _A = 2 years	Range _B = 169 years

Measures of Dispersion

We will study these five measures of dispersion

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

Normal Distribution



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?

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

- ▶ 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:

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

- ▶ 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: N vs. (n-1)

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

- 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, \bar{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, $s = \sqrt{\text{variance}}$.

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

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.

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).

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

Coefficient of Variation (CV)

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

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.

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.

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.

$Z = \text{Score of value} - \text{mean of scores} \div \text{standard deviation}.$

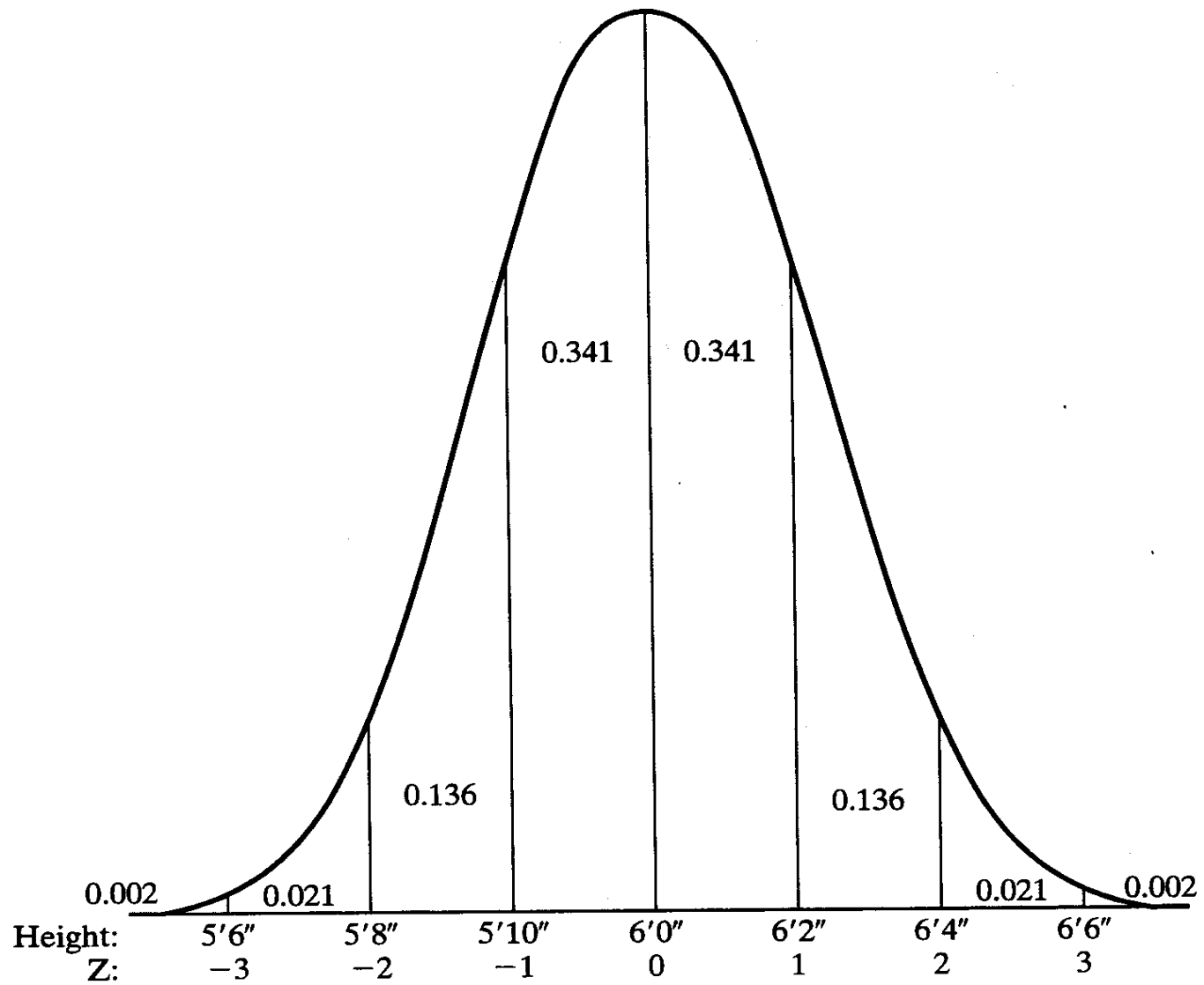
Standard Normal Scores

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.



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 $\bar{X} = 0$ 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

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,

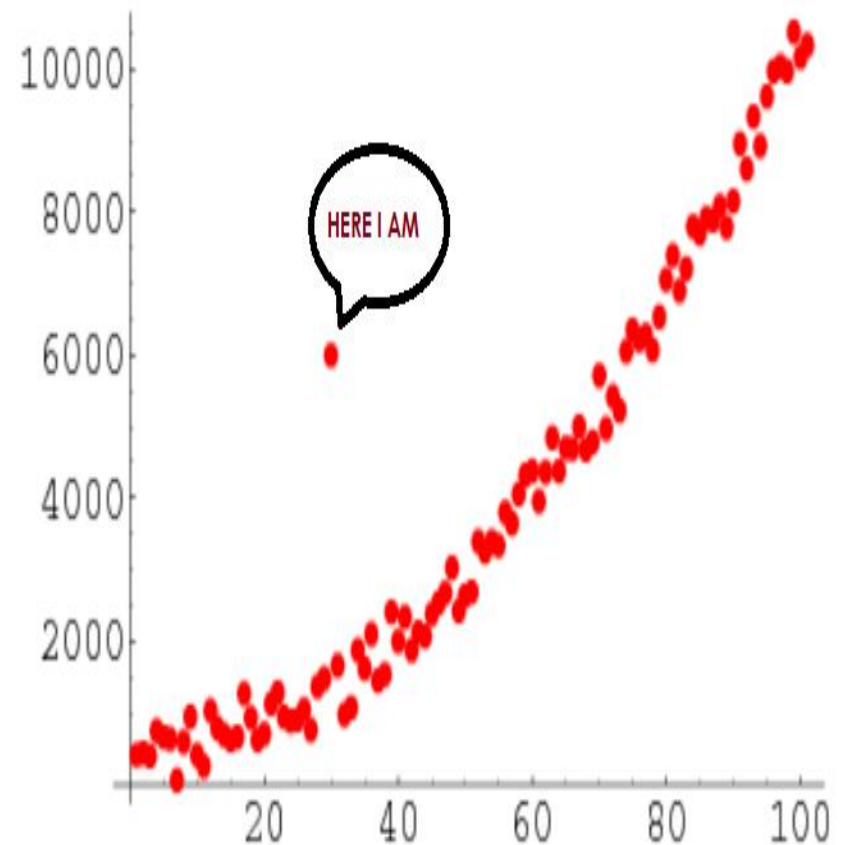
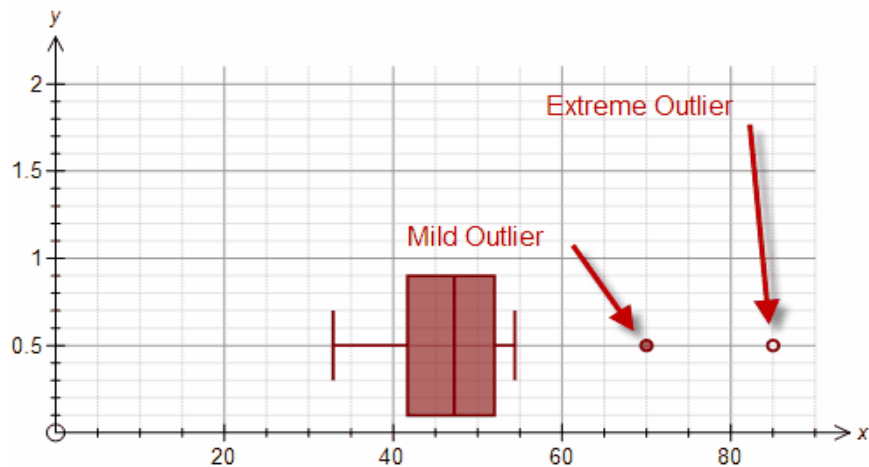
- 95% of the data will be between -1.96 and +1.96.
- 99.7% of the data will fall between -3 and +3.
- 99.99% of the data will fall between -4 and +4.

Major reasons for using Index (descriptive statistics)

Understanding the data.

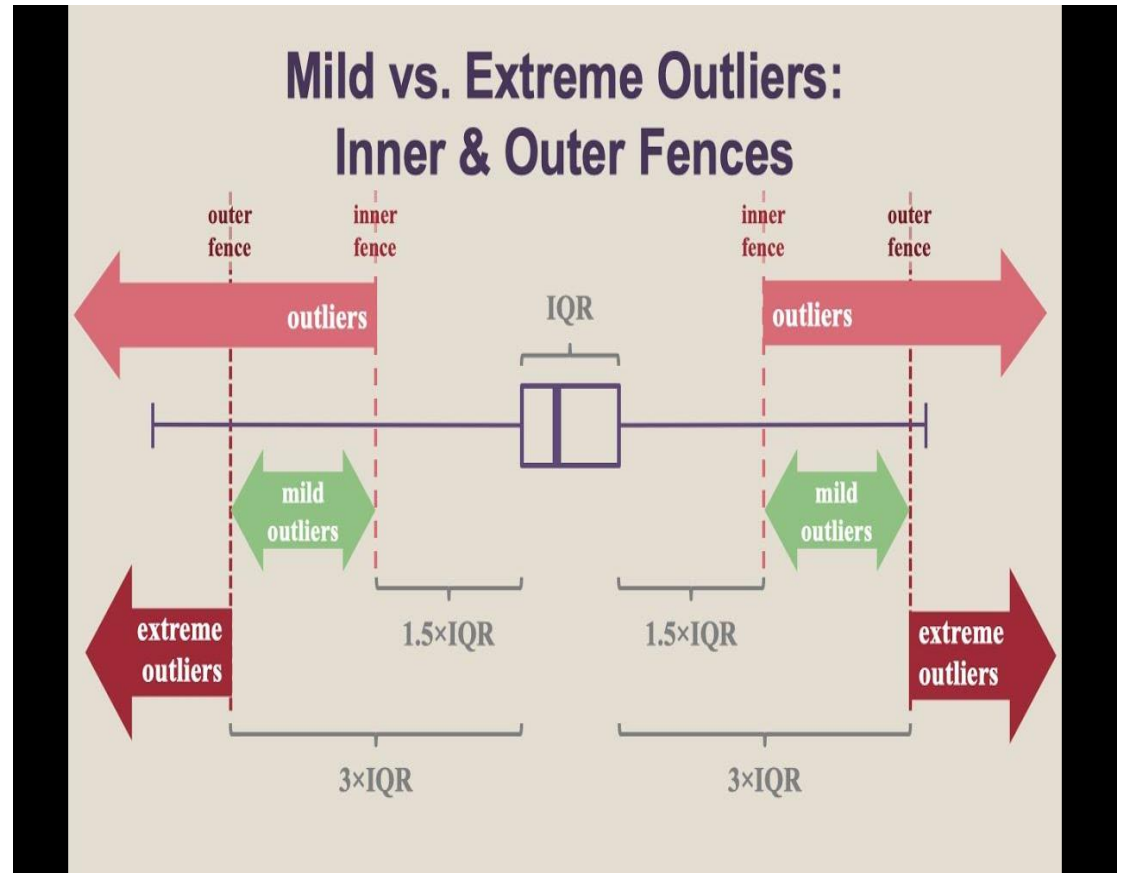
Evaluating outliers. Outliers are often identified in relation to the value of a distribution's IQR.

Describe the research sample.



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 than three times the IQR below Q1 or above Q3.



A photograph of a hand holding a small brown bird, with several other similar birds in flight in the background. The scene is set against a warm, blurred background with bokeh light effects. A semi-transparent dark horizontal band is overlaid across the middle of the image.

THE END OF UNIT 2