FREQUENCY DISTRIBUTIONS, CENTRAL TENDENCY, & VARIABILITY

DESCRIBING DATA

- **Descriptive Statistics:** Techniques for organizing data in ways that make it easy to understand

- We will cover the following descriptive statistics:
  - Frequency Distributions
  - Measures of Central Tendency
  - Measures of Variability

FREQUENCY DISTRIBUTIONS

- A **frequency distribution** is a representation of the number of times each value on a variable occurred in a data set.

- Frequency distributions can be represented using **tables** and **graphs**.
  - One type of table for representing data is an **ungrouped frequency distribution table.**

UNGROUPED FREQUENCY DISTRIBUTION TABLE

An **ungrouped frequency distribution table** is a table made up of the following columns:

- First column: the name of the variable & a list of **all possible** values listed from highest to lowest
- Second column: the frequency of each value in a set of data
- Third column: the cumulative frequency (f_c), which is the frequency of a given value *or* lower than the given value
- Fourth column: percentages column, which is the frequencies transformed into percentages (% = f/N x 100)
- Fifth column: the cumulative percentage, which is the cumulative frequencies transformed into percentages ($\%_c = f_c/N \times 100$)

Number of Children in Family	Frequency (f)	Cumulative Frequency (f_c)	Percentage (%)	Cumulative Percentage ($\%_c$)

UNGROUPED FREQUENCY DISTRIBUTION TABLE

Example: You ask 31 people how many children they have in their family, and you get the following data:

1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 4, 4, 6

Let's fill in the table:

Number of Children in Family	Frequency (f)	Cumulative Frequency (f_c)	Percentage (%)	Cumulative Percentage ($\%_c$)

PRACTICE PROBLEM

A botany professor wants to know how many hours their students spent studying for the last quiz that was given in class. The professor asks the 25 students in their class to report how many hours they spent studying for the last quiz to the nearest hour and collects the following data:

1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 7

Construct an ungrouped frequency table representing the frequency of the students' responses.

Question: What is the cumulative percentage for a score of 5 in the table?

a) 28% b) 100% c) 0% d) 96%

PRACTICE PROBLEM

Solution

Answer: d) 96%

Hours Spent Studying	Frequency (f)	Cumulative Frequency (f_c)	Percentage (%)	Cumulative Percentage ($\%_C$)
7	1	25	4.00	100
6	0	24	0.00	96
5	0	24	0.00	96
4	6	24	24.00	96
3	7	18	28.00	72
2	6	11	24.00	44
1	5	5	20.00	20

GRAPHING FREQUENCY DISTRIBUTIONS

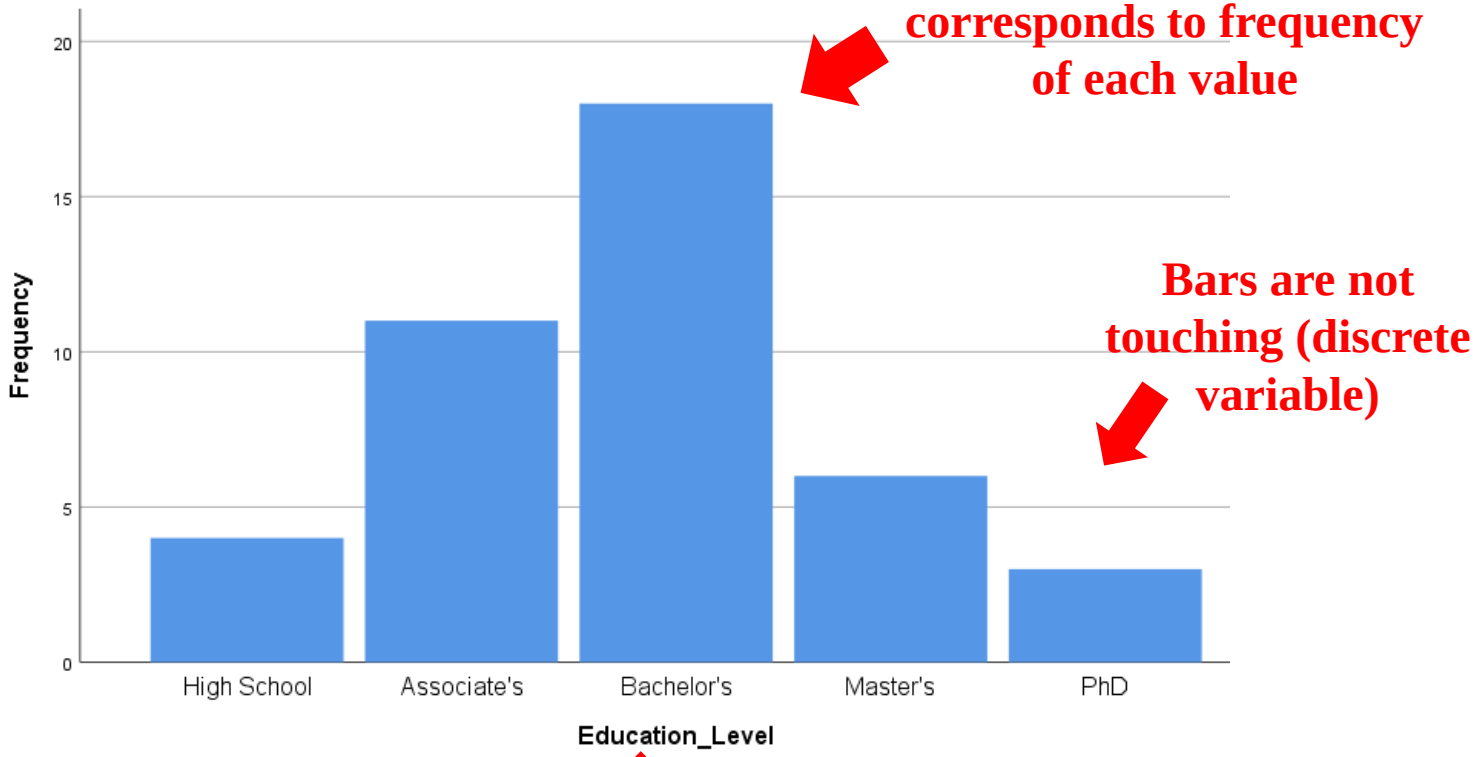
- Frequency distributions can also be represented using graphs.
  - The type of graph used depends on the nature of the variable (discrete vs continuous).

There are three common graphs used:

- Bar graphs
- Histograms
- Frequency polygons

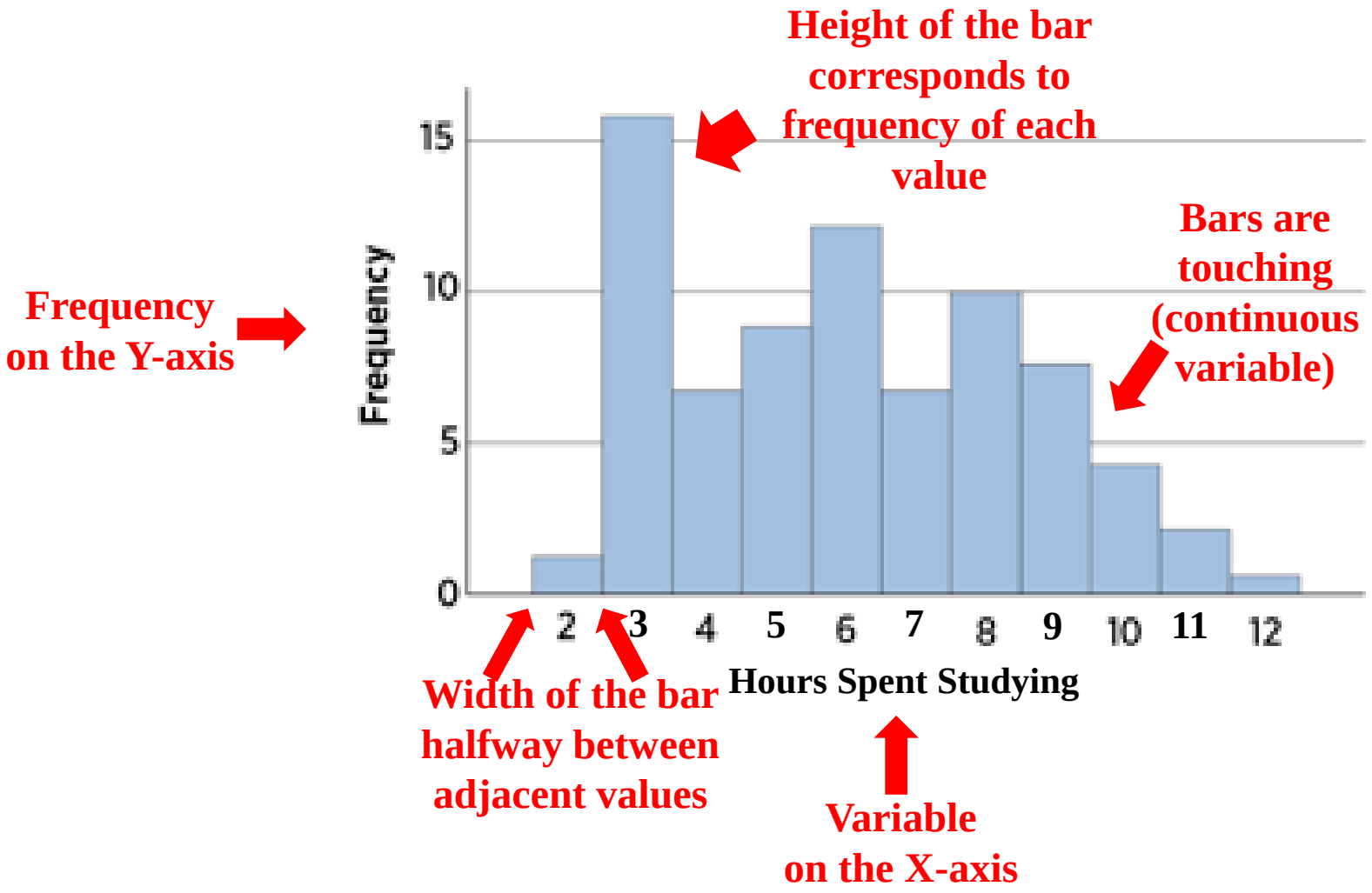
CREATING BAR GRAPHS

Bar graphs are used to represent **discrete variables.**



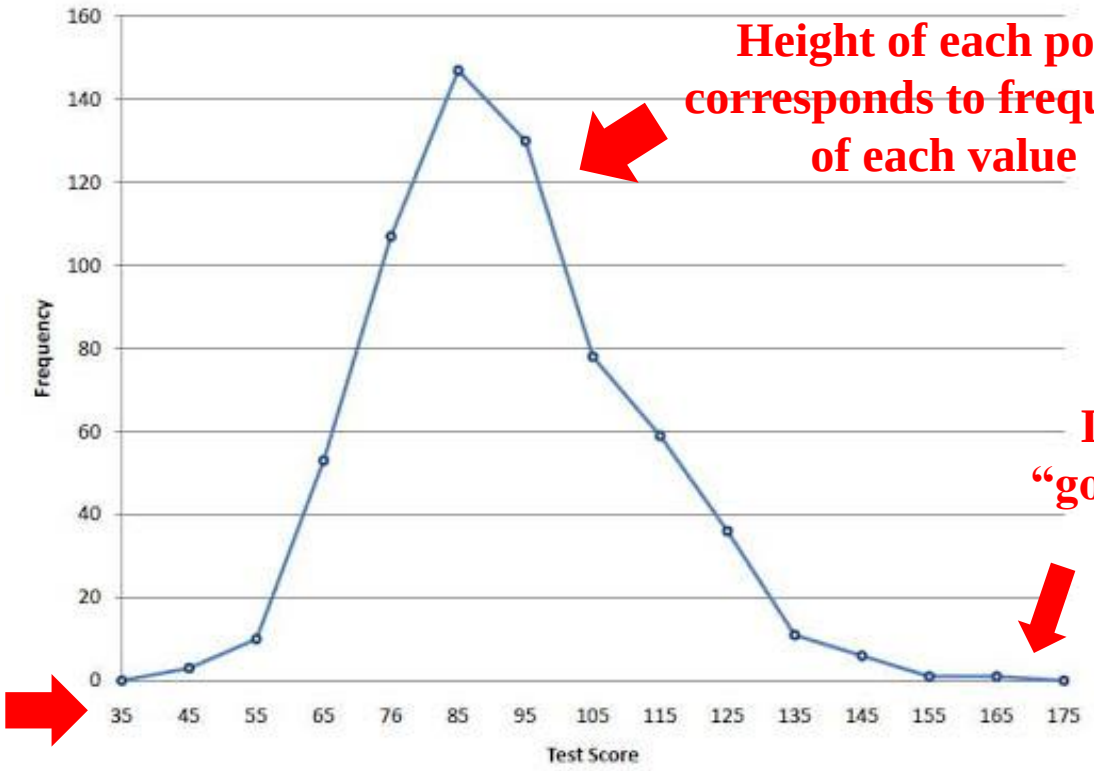
CREATING HISTOGRAMS

Histograms are used to represent **continuous variables**.



CREATING FREQUENCY POLYGONS

Frequency polygons are also used to represent **continuous variables**.



Height of each point corresponds to frequency of each value

Frequency on the Y-axis

Include a line that "goes to zero" at either end of the graph

- **On the x-axis, list each possible value as listed in a frequency table from lowest to highest**

Variable on the X-axis

SHAPES OF FREQUENCY DISTRIBUTIONS

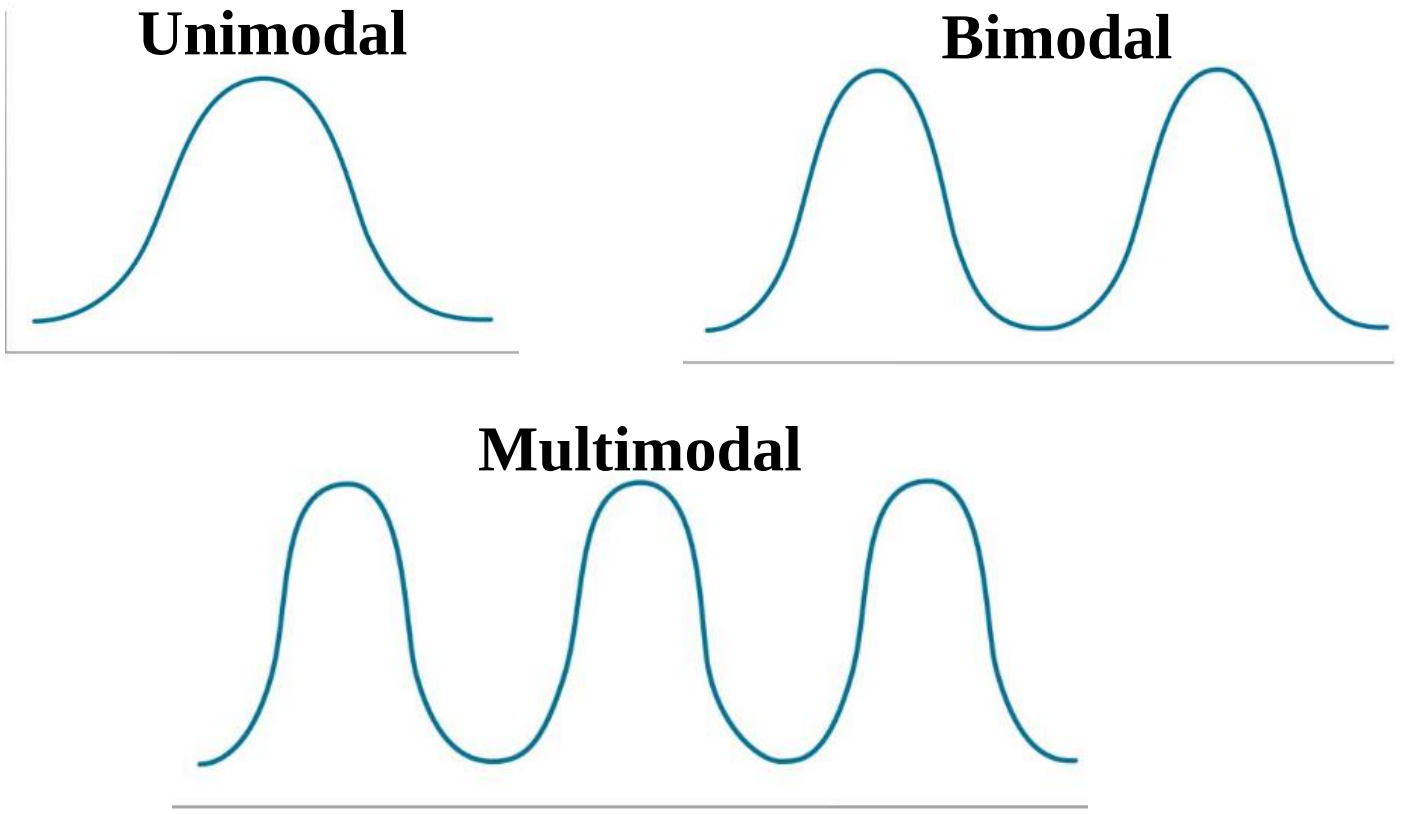
Once a frequency distribution graph has been created, you can analyze its shape.

There are three characteristics about the shape of distributions:

- Modality
- Skewness
- Kurtosis

MODALITY

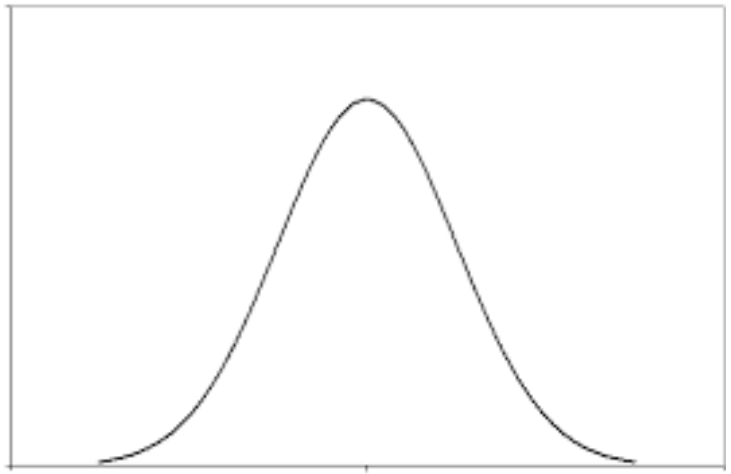
The modality of a frequency distribution graph refers to how many high points (i.e., peaks) it has. Peaks represent values with the highest frequency.



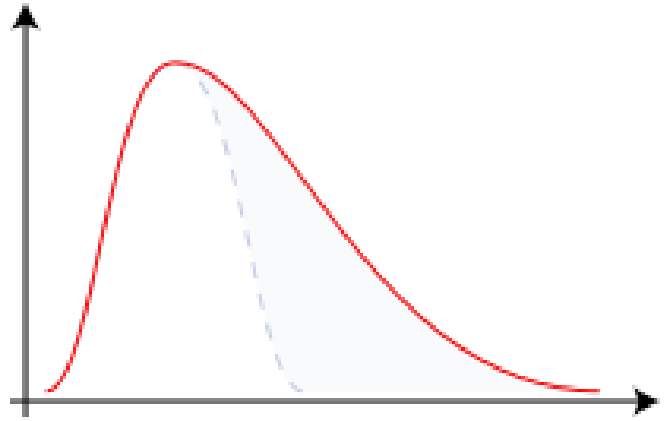
SKEWNESS

Skewness is a measure of how symmetric a frequency distribution is.

Symmetrical distributions - the pattern of frequencies on one half of the distribution is a mirror image of the pattern of frequencies on the other half



Skewed distributions – the majority of scores "pile up" on one side of the distribution, and the rest are spread out on the other side

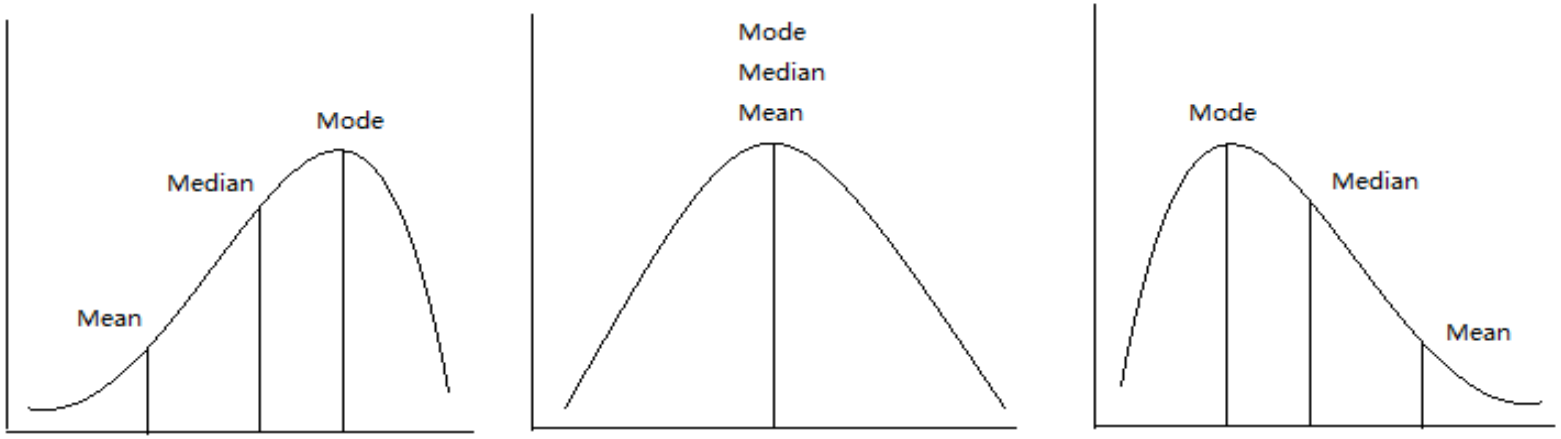


SKEWNESS

The side of the distribution with the *fewer* scores (the "tail") is considered the *direction* of the skew.

Negatively Skewed – fewer scores on the negative side of the distribution

Positively Skewed – fewer scores on the positive side of the distribution



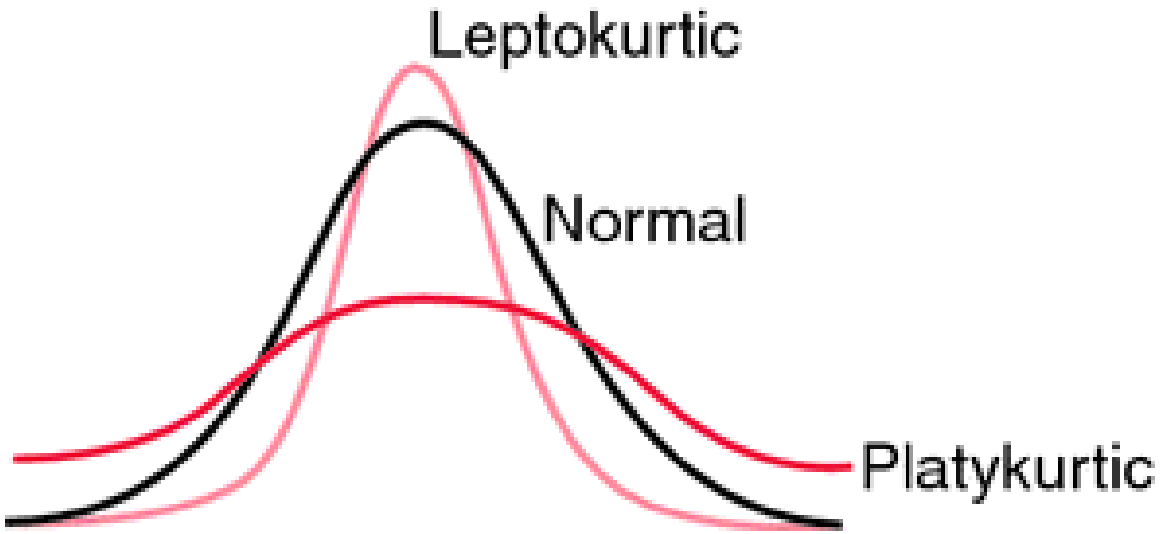
Negatively Skewed | Symmetrical | Positively Skewed

KURTOSIS

The refers to how peaked or flat a frequency distribution is.

Platykurtic – flatter than a normal distribution

Leptokurtic - more peaked than a normal distribution



CENTRAL TENDENCY

A measure of central tendency is a single score that is *representative* of a large data set.

There are three main measures of central tendency:

- Mean
- Median
- Mode

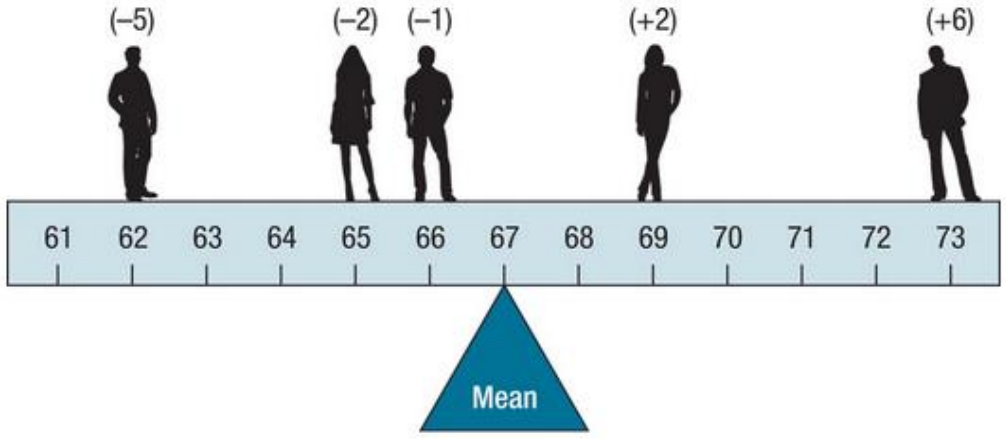
MEAN

The mean is the mathematical average of a set of values. It's calculated by dividing the sum of all scores in a data set by the number of scores:

$$M = \frac{\Sigma X}{n}$$

- M is the symbol for the mean (when you have sample data)
- ΣX is the symbol for the sum of all scores
- n is the number of scores in the sample
  - **Note**: For population-level data, the symbol for the mean is μ

The mean is a *balancing point*. The total distance between the scores above the mean and the mean itself equals the total distance between the scores below the mean and the mean itself.



MEDIAN

The median is the midpoint of a set of scores when they are listed from lowest to highest.

To calculate the median:

1) List scores from lowest to highest
2) Find the middle score
   - When the scores are listed from highest to lowest, the **location** of the median (**not** the median itself) can be found by using the calculation: (n + 1)/2

If you have two scores in the middle, the median is the average of the two scores.

MODE

The mode is the most frequently occurring score (relative to the scores around it)

- There can be more than one mode

COMPARING THE MEAN, MEDIAN, & MODE

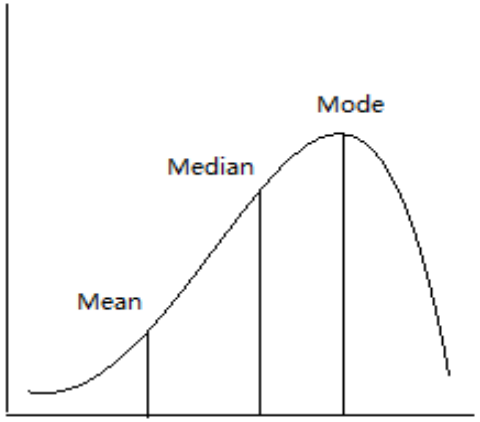
- The **mean** is the most commonly used in psychological research
  - Strength: takes *all* scores into consideration
  - Weakness: affected by extreme scores

- When there are extreme scores (i.e., the distribution is skewed) the **median** is a better representation of the most typical score.
  - Strength: Not as affected by extreme scores as the mean
  - Weakness: Not all scores contribute *equally* (i.e., you could change some of the score's values without it affecting the median)

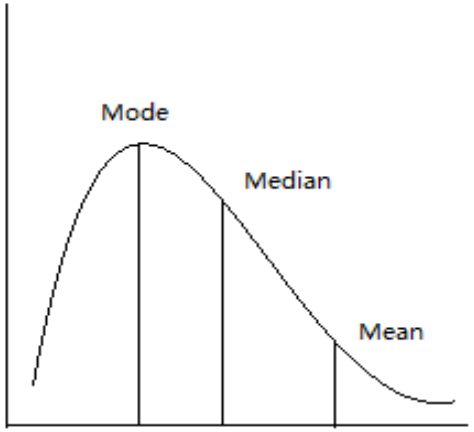
- The **mode** is the only measure of central tendency that can be used for nominal data.
  - Strength: Can be used as a representative value for nominal data
  - Weakness: Only descriptive of one value in the data set

HOW THE MEAN IS AFFECTED BY EXTREME SCORES

The mean is pulled **towards** the extreme scores and **away** from where the majority of scores are piling up.



Negatively Skewed



Positively Skewed

SUMMARY OF CENTRAL TENDENCY MEASURES

Measure	Definition	When to use
Mean	Sum of the scores divided by the number of scores	• Interval or ratio variables • Most commonly used in psychological research
Median	The middle score when scores are listed from lowest to highest	• When there are extreme scores (i.e., a skewed distribution) • Interval, ratio, or ordinal variables
Mode	The value(s) with the greatest frequency in a set of scores	• Nominal variables

PRACTICE PROBLEM

1) A school psychologist measures aggression levels in seven elementary school children. They obtain the following data: 98, 100, 99, 97, 102, 97, 101

Q: Which measure of central tendency is best to use in this case?

a) Mode b) Median c) Mean

2) A sensory psychologist gives participants several candies to try and asks them to classify the candy as either bitter (1), salty (2), savory (3), sour (4), or sweet. They obtain the following data: 1, 2, 3, 4, 5, 5, 5, 3, 2, 3, 4, 5, 1, 5, 4, 5

Q: Which measure of central tendency is best to use in this case?

a) Mode b) Median c) Mean

PRACTICE PROBLEM

Solutions

1) Answer: c) Mean

2) Answer: a) Mode

VARIABILITY

The central, or typical, score in a set of data is only one aspect of a data set. Researchers are also interested in how scores vary from one another.

Variability refers to how spread out the scores in a set of data are.

There are four typical measures of variability:

- Range
- Interquartile Range (IQR)
- Variance
- Standard Deviation

THE RANGE

The range is the difference between the highest score and the lowest score in a set of data.

$$\text{Range} = X_{\text{Highest}} - X_{\text{Lowest}}$$

Example: In a set of data, the oldest person who responded was 50, and the youngest was 16.

- Range = 50 – 16 = **34**

Weakness: The range **only** expresses the difference between the **most extreme** scores in a data set.

THE INTERQUARTILE RANGE

The interquartile range (IQR) is a measure of range for only the middle half of a group of scores.

Steps for calculating the IQR:

1. Arrange scores from smallest to largest
2. Find the median. This is the second quartile (Q2).
3. Find the first quartile (Q1). This is the median for the scores below Q2.
4. Find the third quartile (Q3). This is the median for the scores above Q2.

The interquartile range is the range between Q1 and Q3.

$$IQR = Q3 - Q1$$

Example: Find the IQR for: 12, 13, 13, 14, 15, 17, 17, 18, 20

VARIANCE & STANDARD DEVIATION

- **Variance** and **standard deviation** are commonly used measures of variability in psychology. They measure how much, on average, scores in a data set deviate from the mean.

- **Variance**: average squared deviation of scores from the mean

- **Standard Deviation**: average non-squared deviation of scores from the mean

CALCULATING VARIANCE & STANDARD DEVIATION

To calculate the standard deviation:

(1) Find the average for a group of scores.

(2) Calculate a **deviation score** for each score by subtracting the mean from the score.

Calculate the standard deviation for the following set of scores:
3, 5, 7, 9, 11

X	(X – M)
$M = \frac{\Sigma X}{n}$	

What happens when you sum the deviation scores?

CALCULATING VARIANCE & STANDARD DEVIATION

(3) Square each of the deviation scores.

(4) Find the **sum of the squared deviation scores (SS)**.

X	$(X - M)$	$(X - M)^2$
3	-4	16
5	-2	4
7	0	0
9	2	4
11	4	16
$M = 7$		$SS = \Sigma(X - M)^2$

CALCULATING VARIANCE & STANDARD DEVIATION

If you have **population** data, the next steps are:

(5) Divide the SS by the population size. This gives you the **population variance**.

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

(6) Take the square root of the variance. This gives you the **population standard deviation.**

$$\sigma = \sqrt{\frac{SS}{N}}$$

If you have **sample** data, the next steps are:

(5) Divide the SS by the sample size minus one. This gives you the **estimate of the population variance**.

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

(6) Take the square root of the variance. This gives you the **estimate of the population standard deviation.**

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

DEGREES OF FREEDOM

When we have sample data, why do we divide by n – 1 instead of n?

- The denominator, n – 1, is called the **degrees of freedom**.

Remember the role of **inferential statistics** – we are limited by practical limitations to only having sample data, but we want to *draw conclusions* about a larger population.

- What effect does dividing by n – 1 instead of n have?

PRACTICE PROBLEM

The American Association of Organic Apple Growers measured the number of acres in five family farms. The results were:

23, 8, 22, 10, 32

For this set of **sample** data, calculate:

- Range
- IQR
- Estimate of the population variance
- Estimate of the population standard deviation

PRACTICE PROBLEM

Q: What is the range? a) 24 b) 32 c) 21

Q: What is the IQR? a) 18.5 b) 13 c) 5.5

Q: What is the variance? a) 9.95 b) 396 c) 99

Q: What is the standard deviation? a) 9.95 b) 396 c) 99

PRACTICE PROBLEM

Solutions

What is the range? **a) 24** b) 32 c) 21

What is the IQR? **a) 18.5** b) 13 c) 5.5

What is the variance? a) 9.95 b) 396 **c) 99**

What is the standard deviation? **a) 9.95** b) 396 c) 99

REVIEW OF NOTATION

	Sample Data	Population Data
Sample Size	n	N
Mean	M	μ
Variance	s^2	σ^2
Standard Deviation	s, SD	σ

DESCRIPTIVE STATISTICS

- Descriptive statistics for our class data!

WEEK 1 REMINDERS

- Homework 1
  - Based on the material covered in lecture during week 1
  - Due Monday, June 28th, at 11:59pm
  - Access it via the "Assignments" tab on Canvas

- Jamovi HW1
  - Based on the material covered in lab during week 1
  - Due Thursday, July 1st, at 11:59pm
  - Access it on Canvas (more information will be provided by your lab instructor!)
  - Lab 1 meets this Friday at 10am (Zoom link on Canvas)