Introduction to Statistics

Statistics, Science, and Observations

- Research in behavioral sciences involves gathering information
  - E.g., whether college students learn better by reading material on printed pages or on a computer screen
  - Survey about students' study habits and their academic performance

Statistics, Science, and Observations

- Research in behavioral sciences involves gathering information
  - E.g., whether college students learn better by reading material on printed pages or on a computer screen
  - Survey about students' study habits and their academic performance
- Analyze (e.g., organize and summarize)
  - E.g., printed page, average score = 26; computer screen, average score = 22
- Interpret

Statistics, Science, and Observations

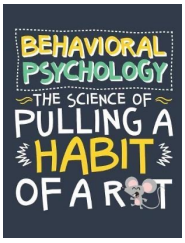
- **Statistics**: A set of mathematical procedures for *organizing*, *summarizing*, and *interpreting* information
- Ensure that the information or observations are presented and interpreted in an accurate and informative way
  - Bring order out of chaos
- Provide a set of standardized techniques that are recognized and understood throughout the scientific community

Statistics, Science, and Observations

- Research in behavioral sciences typically begins with a general question about a specific group (or groups) of individuals
  - E.g., what factors are associated with academic dishonesty among **college students**
  - E.g., the amount of time spent in the bathroom for **men** compared to **women**
- **Population**: The set of *all* the individuals of interest in a particular study

Statistics, Science, and Observations

- A population can be quite large
  - E.g., all the men and women on Earth
- Populations can obviously vary in size
  - Depending on how the investigator defines the population
- Need not consist of people
  - A population of rats, corporations, parts produced in a factory …
  - Anything an investigator wants to study



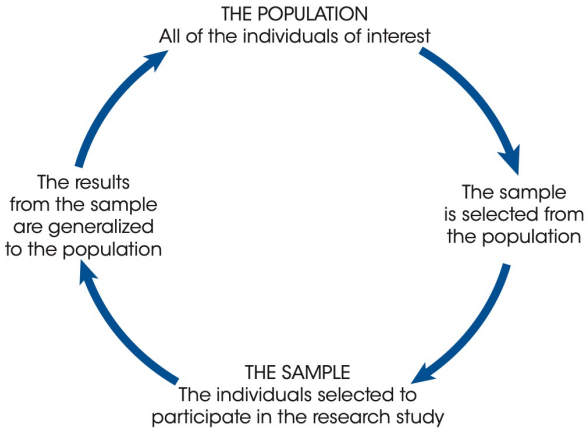
Statistics, Science, and Observations

- In practice, populations are typically very large
- Impossible to examine every individual in the population of interest
- Select a smaller, more manageable group from the population
- Limit the studies to the individuals in the selected group
- **Sample**: A set of individuals selected from a population, usually intended to represent the population in a research study

Statistics, Science, and Observations

- Samples can vary in size
  - E.g., a sample of 10 students in a graduate program
  - E.g., a sample of 10,000 people who got Moderna's vaccine
- A sample being selected from a population
  - Half of the full relationship
- Goal: generalize the results back to the entire population

Statistics, Science, and Observations



Statistics, Science, and Observations

- Typically, researchers are interested in specific characteristics of individuals in the population/sample
  - E.g., height, weight, IQ, attention span…
- Or the outside factors that may influence the individuals
  - E.g., the influence of the weather on people's mood
  - As the weather changes, do people's moods also *change*?
- **Variable**: A characteristic or condition that *changes* or has *different values* for different individuals

Statistics, Science, and Observations

- Variables can be characteristics that differ from one individual to another
  - E.g., height, weight, gender, personality…
- Variables can be environmental conditions that change
  - E.g., temperature, time of day, the size of the room in which the study is being conducted …
- To demonstrate changes in variables, make measurements of the variables being examined.
- **Data** (plural): measurements or observations
- **Datum** (singular): a single measurement or observation
  - A **score** or **raw score**

Statistics, Science, and Observations

- When describing data it is necessary to distinguish whether the data come from a population or a sample
- **Parameter**: a value, usually a numerical value, that describes a population
  - usually derived from measurements of the individuals in the population
- **Statistic**: a value, usually a numerical value, that describes a sample
  - usually derived from measurements of the individuals in the sample
- Every population parameter has a corresponding sample statistic
- Most research studies involve using statistics from samples as the basis for answering questions about population parameters

Statistics, Science, and Observations

- Researchers have developed a variety of different statistical procedures to organize and interpret data
- Two categories
- **Descriptive statistics**: statistical procedures used to summarize, organize, and simplify data
  - Raw scores → organize (table or graph) or summarize (computing an average)
- **Inferential statistics**: techniques that allow us to study samples and then make *generalizations* about the populations from which they were selected

Statistics, Science, and Observations

- Researchers use sample statistics as the basis for drawing conclusions about population parameters
- One problem
  - A sample provides only limited information about the population
- Samples are generally *representative* of their populations
- A sample is not expected to give a perfectly accurate picture of the whole population
  - Some discrepancy
- **Sampling error**: the naturally occurring discrepancy, or error, that exists between a sample statistic and the corresponding population parameter

Statistics, Science, and Observations

- Population of 100 students at College XYZ (variable: age; average = 24.17)

27	18	20	23	18	27	22	20	23	28
26	18	22	18	29	19	27	22	29	24
30	25	27	29	26	28	19	29	25	26
23	24	22	26	28	21	18	18	18	26
26	30	27	22	19	27	21	21	23	26
21	22	30	19	29	28	27	18	29	27
22	28	19	21	29	25	29	18	22	27
29	21	26	20	28	25	23	20	18	27
30	30	30	28	24	21	22	25	20	23
19	18	20	23	29	27	26	30	29	26

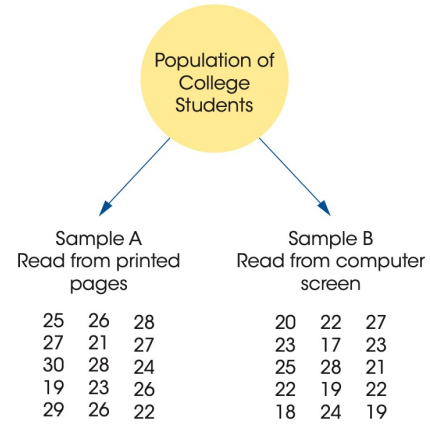
Statistics, Science, and Observations

- Statistics in the Context of Research
  - Do college students learn better by studying text on printed pages or on a computer screen?

Statistics, Science, and Observations

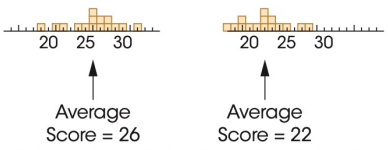
Step 1
Experiment:
Compare two studying methods

Data
Test scores for the students in each sample



Statistics, Science, and Observations

Step 2
Descriptive statistics:
Organize and simplify



Statistics, Science, and Observations

Step 3
Inferential statistics:
Interpret results

The sample data show a 4-point difference between the two methods of studying. However, there are two ways to interpret the results.

1. There actually is no difference between the two studying methods, and the sample difference is due to chance (sampling error).
2. There really is a difference between the two methods, and the sample data accurately reflect this difference.

The goal of inferential statistics is to help researchers decide between the two interpretations.

Statistics, Science, and Observations

- **Statistics**: A set of mathematical procedures for *organizing*, *summarizing*, and *interpreting* information
- **Population**: The set of *all* the individuals of interest in a particular study
- **Sample**: A set of individuals selected from a population, usually intended to represent the population in a research study
- **Variable**: A characteristic or condition that *changes* or has *different values* for different individuals
- **Data** (plural): measurements or observations
- **Parameter**: a value, usually a numerical value, that describes a population
- **Statistic**: a value, usually a numerical value, that describes a sample
- **Descriptive statistics**: statistical procedures used to summarize, organize, and simplify data
- **Inferential statistics**: techniques that allow us to study samples and then make *generalizations* about the populations from which they were selected

Data and Levels of Measurement

Variables and Data

- Variables can be characteristics that differ from one individual to another
  - E.g., height, weight, gender, personality…
- Variables can be environmental conditions that change
  - E.g., temperature, time of day, the size of the room in which the study is being conducted …
- To demonstrate changes in variables, make measurements of the variables being examined.
- **Data** (plural): measurements or observations
- **Datum** (singular): a single measurement or observation
  - A **score** or **raw score**

Data Types

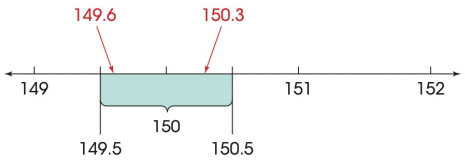
- A. Hair color: black, brown, black, black, red, brown, blonde, gray…
- B. The car one drives: BMW, Toyota, Subaru, Ford, BMW, Ford…
- C. Age: 25, 35, 20, 18, 20, 21, 30, 20, 22…
- D. Body temperature: 36.5, 36.4, 37.1, 36.7, 37, 36.6…
- E. Country of residence: Canada, Canada, US, Mexico, US, Brazil…
- F. Exam score: 85, 75, 75, 80, 95, 90, 80, 90…

Data Types

- Qualitative data
  - Observation represents a class or category
  - Categorical data
  - Described by words or letters
- Quantitative data
  - Result of counting or measuring
  - Always numbers
  - Often preferred by researchers
    - Mathematical analysis (e.g., average of exam scores)
    - Not make sense to find an average hair color or blood type

Data Types

- Quantitative data: result of **counting** or **measuring**
- Discrete: counting
  - Countable with integers
  - E.g., **number of** students, cars, phone calls…
- Continuous: measuring
  - Consist of an infinite number of possible values on a scale
  - May include fractions, decimals, or irrational numbers



Measurement

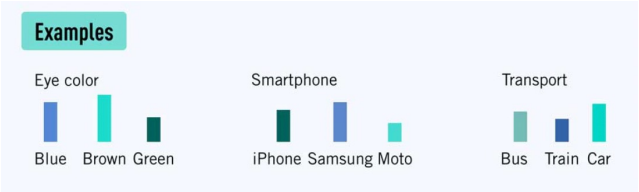
- Analyze the spending habits of people living in Toronto
- Survey 500 people
- **Income**, age, how much they spend on various products…
- What kinds of data relating to people's income can you gather?
  - Provide an exact figure ($109,000, $35,000, $21,000)
  - Select from a variety of ranges: (a) 10-19k, (b) 20-29k, (c) 30-39k…
  - Categorize income as: "high", "medium", "low"
- Levels vary in precision

Levels of Measurement

- The way a set of data is measured
  - How precisely
- Determine the type of statistical analysis you can carry out
  - Certain statistical tests can only be performed where more precise levels of measurement have been used
- Four levels/scales of measurement (or 4 different types of data)
  - Least complex and precise → most
    - Nominal
    - Ordinal
    - Interval
    - Ratio

Nominal Scale

- "Nominal" from Latin for "name"
- Qualitative (categorical) data
  - Consist of names, labels, or categories



Nominal Scale

- "Nominal" from Latin for "name"
- A type of qualitative (categorical) data
  - Consist of names, labels, or categories
- **Data cannot be meaningfully arranged in order**
  - No one category is greater/better than or "worth more" than another
  - E.g., trying to order people according to their favorite food does not make any sense. Putting vegie pizza first and vegan sushi second is not meaningful.

Nominal Scale

- May be represented by numbers
  - "Number labels", no numeric meaning
- Example
  - Collecting data on people's hair color
  - 1: brown hair; 2: blonde hair; 3: black hair
  - These numbers do not represent any kind of value or hierarchy
    - Black hair represented by 3 is not greater/better than brown hair represented by 1, and vice versa

Ordinal Scale

- A type of qualitative (categorical) data
- The categories are **ordered** on some kind of hierarchical scale
  - E.g., high to low
- Category + Order

Ordinal Scale

- May be represented by numbers
  - "Number labels", no numeric meaning
- Example
  - Collecting data on people's economic status
  - 1: poor; 2: middle income; 3: wealthy
  - 3: poor; 2: middle income; 1: wealthy

Interval Scale

- A type of quantitative data
- Differences or distances (intervals) between the scores are meaningful
- Category + Order + Intervals
- Example:
  - Temperature in Fahrenheit or Celsius (-20, -10, 0, +10, +20, etc.)
    - 100 is 20 degrees hotter than 80

Interval Scale

- A zero point may be lacking or it may be arbitrary
- A measure of zero does not denote the absence of something
  - Zero degrees in both Celsius and Fahrenheit temperature scales do NOT mean the absence of heat
  - There can be temperatures below zero degrees
- 70 degrees is “twice as hot” as 35 degrees?
  - **No**
  - 30 degrees is -1 times as hot as -30 degrees?

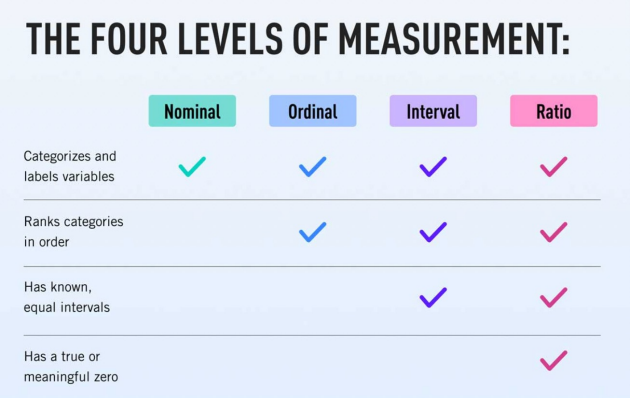
Ratio Scale

- An interval scale with a true zero
- True zero point: complete absence of data being measured
  - Can never have a negative value
  - You cannot be -10 years old or weigh -160 pounds!
- Category + Order + Intervals + True 0
- Examples
  - Temperature in Kelvin (0, +10, +20, +30, +40, etc.)
  - Height (5ft. 8in., 5ft. 9in., 5ft. 10in., 5ft. 11in., 6ft. 0in. etc.)
  - Price of goods ($0, $5, $10, $15, $20, $30, etc.)
  - Age in years (from zero to 100+)

Ratio Scale

- The existence of true zero simply means that the measurement scale you are using has a definitive starting point of zero, i.e. you could reach zero in theory, even if not in practice.
  - E.g., if you're measuring the heights of a group of adults, you probably won't obtain many measurements below 5 feet.

Levels of Measurement



Levels of Measurement

- Determine the type of statistical analysis you can carry out
  - Certain statistical tests can only be performed where more precise levels of measurement have been used
- The vast majority of the statistical techniques covered in this course are designed for numerical scores from interval or ratio scales
  - E.g., measure IQ scores for a group of students: add the scores together to find a total and then calculate the average score for the group
  - E.g., measure the academic major for a group of students: What is the total for three psychology majors plus an English major plus two chemistry majors?

Levels of Measurement

The teacher in a communications class asks students to identify their favorite reality television show. The different television shows make up a _____ scale of measurement.

A. Nominal

B. Ordinal

C. Interval

D. Ratio

Levels of Measurement

The teacher in a communications class asks students to report how many hours they spent watching reality television shows in the past week. The different hours make up a _____ scale of measurement.

A. Nominal

B. Ordinal

C. Interval

D. Ratio

Levels of Measurement

- A researcher obtains of measurements of height for a group of 10 people: 170cm, 160cm, 180cm, 190cm, 160cm, 170cm, 170cm, 170cm, 180cm, 150cm
- A _____ scale of measurement
  - Ratio
- The researcher converts the initial measurement into a new scale by calculating the difference between each person's actual height and the average height for this group (average = 170cm)
  - 0cm, -10cm, 10cm, 20cm, -10cm, 0cm, 0cm, 0cm, 10cm, -20cm
- A _____ scale of measurement
  - Interval
  - Zero no longer indicates an absence of height
  - 20cm is not twice as tall as 10cm

Research Methods and Ethics

Individual Variables

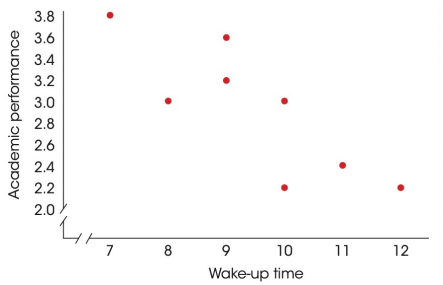
- Simply to describe individual variables as they exist naturally (descriptive research)
  - E.g., to describe the eating, sleeping, and study habits of a group of college students

Relationships between Variables

- Most research is intended to examine relationships between two or more variables
  - Examples:
  - The amount of violence in the video games played by children and the amount of aggressive behavior they display
  - The quality of breakfast and academic performance for elementary school children
  - The number of hours of sleep and GPA for college students
- To establish the existence of a relationship, researchers must make observations/measurements of the two variables

Correlational Method

- Observe the two variables as they exist naturally for a set of individuals
  - Simply measure the two variables for each individual
- Look for consistent patterns in the data to provide evidence for a relationship between variables



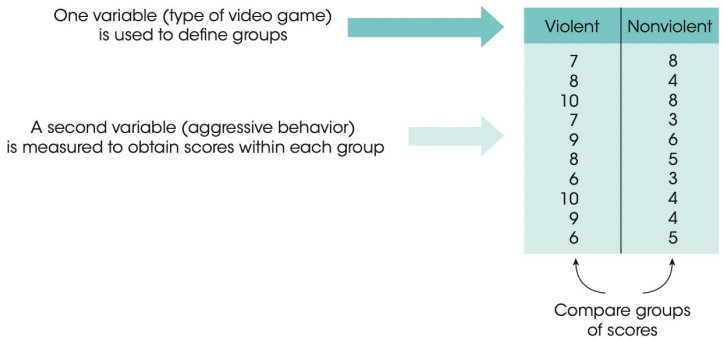
Correlational Method

- Can demonstrate the existence of a relationship between two variables
- Do NOT provide an explanation for the relationship
  - Cannot demonstrate a cause-and-effect relationship
- Example: the amount of violence in the video games played by children and the amount of aggressive behavior they display
  - More violence in the video games causes more aggressive behavior?
  - More aggressive children tend to play more violent video games?

Experimental Method

- The comparison of two or more groups of scores
- Use one of the variables to define the groups, and then measure the second variable to obtain scores for each group
- Example: the amount of violence in the video games played by children and the amount of aggressive behavior they display
  - Randomly divide a sample of 10-year-old boys into two groups
  - One group: violent game; the other: nonviolent game
  - After the game-playing session, the children go on to a free play period and were monitored for aggressive behaviors
  - Compare the scores for the violent-game group with those for the nonviolent-game group

Experimental Method



Experimental Method

- Goal: To demonstrate a cause-and-effect relationship between two variables
  - Changing the value of one variable causes changes to occur in the second variable
- Two characteristics
  - **Manipulation**: manipulate one variable by changing its value from one level to another; a second variable is observed/measured to determine whether the manipulation causes changes to occur.
  - **Control**: exercise control over the research situation to ensure that other, extraneous variables do not influence the relationship being examined
- In more complex experiments, a researcher may systematically manipulate more than one variable and may observe more than one variable.

Experimental Method

- **Independent/explanatory variable**: the variable that is manipulated by the researchers
- **Dependent/response variable**: the variable that is observed to assess the effect of the treatment
- **Treatment**: the different values of the independent/explanatory variable

Experimental Method

- **Lurking variables**: additional variables that can cloud a study and must be controlled
  - Participant variables: characteristics such as age, gender, and intelligence that vary from one individual to another
    - Participant variables do not differ from one group to another
    - Violence in video games and aggressive behavior
      - Primarily female in the nonviolent-game group while primarily male in the violent-game group?
  - Environmental variables: characteristics of the environment such as lighting, time of day, and weather conditions
    - Individuals in treatment A are tested in the same environment as the individuals in treatment B
    - Violence in video games and aggressive behavior
      - individuals in the nonviolent condition were all tested in the morning and the individuals in the violent condition were all tested in the evening?

Experimental Method

- Control lurking variables
  - **Random assignment**: each participant has an equal chance of being assigned to each of the treatment conditions
    - To distribute the participant characteristics evenly between the two groups so that neither group is noticeably smarter (or older, or faster) than the other
    - The only difference between groups is the one imposed by the researcher
  - Matching: to ensure equivalent groups or equivalent environments
    - E.g., every group has exactly 60% females and 40% males
  - Holding variables constant
    - E.g., recruit only 10-year-old boys as participants (holding age and gender constant)

Experimental Method

- In the **experimental method**, one variable is manipulated while another variable is observed and measured. To establish a cause-and-effect relationship between the two variables, an experiment attempts to control all other variables to prevent them from influencing the results.

Learning Check

Stephens, Atkins, and Kingston (2009) found that participants were able to tolerate more pain when they shouted their favorite swear words over and over than when they shouted neutral words. For this study, what is the independent variable?

a. the amount of pain tolerated

b. the participants who shouted swear words

c. the participants who shouted neutral words

d. the kind of word shouted by the participants

Learning Check

Weinstein, McDermott, and Roediger (2010) conducted an experiment to evaluate the effectiveness of different study strategies. One part of the study asked students to prepare for a test by reading a passage.

In one condition, students generated and answered questions after reading the passage. In a second condition, students simply read the passage a second time. All students were then given a test on the passage material and the researchers recorded the number of correct answers.

- **Identify the dependent variable for this study.**

Learning Check

Weinstein, McDermott, and Roediger (2010) conducted an experiment to evaluate the effectiveness of different study strategies. One part of the study asked students to prepare for a test by reading a passage.

In one condition, students generated and answered questions after reading the passage. In a second condition, students simply read the passage a second time. All students were then given a test on the passage material and the researchers recorded the number of correct answers.

- **Is the dependent variable discrete or continuous?**

Data Types

- Quantitative data: result of **counting** or **measuring**
- Discrete: counting
  - Countable with integers
  - E.g., **number of** students, cars, phone calls…
- Continuous: measuring
  - Consist of an infinite number of possible values on a scale
  - May include fractions, decimals, or irrational numbers

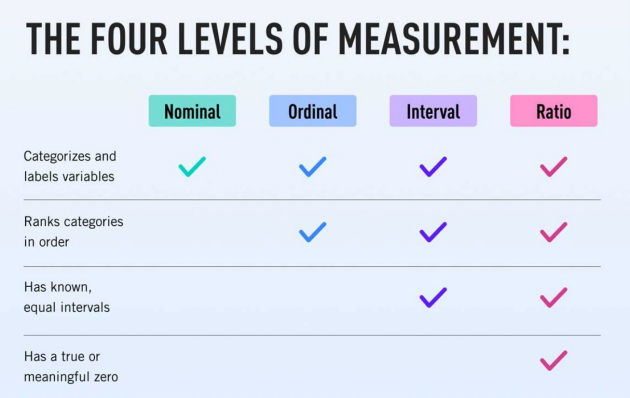
Learning Check

Weinstein, McDermott, and Roediger (2010) conducted an experiment to evaluate the effectiveness of different study strategies. One part of the study asked students to prepare for a test by reading a passage.

In one condition, students generated and answered questions after reading the passage. In a second condition, students simply read the passage a second time. All students were then given a test on the passage material and the researchers recorded the number of correct answers.

- **What scale of measurement (nominal, ordinal, interval, or ratio) is used to measure the dependent variable?**

Levels of Measurement



Learning Check

- The quality of breakfast and academic performance for elementary school children
- How to establish a causal relationship?

Ethics

- Verify that proper methods are being followed
- Be mindful of the safety of their research subjects
  - Risks to participants must be minimized and reasonable with respect to projected benefits.
  - Participants must give **informed consent**. This means that the risks of participation must be clearly explained to the subjects of the study. Subjects must consent in writing, and researchers are required to keep documentation of their consent.
  - Data collected from individuals must be guarded carefully to protect their privacy.
- The misuse of statistics is a bigger problem than most people realize
  - Learning the basic theory of statistics will empower you to analyze statistical studies critically.

Descriptive Statistics

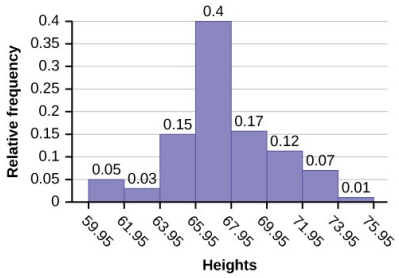
Tables & Graphs

Hmmmm…

- What do we do when we have research questions about a large sample of data, such as various characteristics of graduate students?
  - Carry around the entire collection of data and show it to colleagues

Hmmmm…

- What do we do when we have research questions about a large sample of data, such as various characteristics of graduate students?
  - Summarize the data in an abbreviated fashion through the use of a few tables and/or figures



Tables

- There are a number of different ways to summarize data on a single variable in the form of a table
- Frequency distribution
  - Sorted collection of observations showing the number of times (or **frequency**) each observation occurs

  - Which score occurred most frequently?
  - Which scores were the highest/lowest?
  - Where do most of the scores tend to fall?
  - …

Frequency Distribution

- Example
- Twenty students were asked how many hours they worked per day.
- Their responses, in hours, are as follows:
  - 5 6 3 3 2 4 7 5 2 3 5 6 5 4 4 3 5 2 5 3
- Table lists the different data values in ascending order and their frequencies.

DATA VALUE (*X*)	FREQUENCY (*f*)
2	3
3	5
4	3
5	6
6	2
7	1

Frequency Distribution

- Relative frequency: the frequency for an observed value of the data divided by the total number of data values in the sample
- Relative frequencies can be written as fractions, percents, or decimals

DATA VALUE	FREQUENCY	RELATIVE FREQUENCY
2	3	$\frac{3}{20}$ or 0.15
3	5	$\frac{5}{20}$ or 0.25
4	3	$\frac{3}{20}$ or 0.15
5	6	$\frac{6}{20}$ or 0.30
6	2	$\frac{2}{20}$ or 0.10
7	1	$\frac{1}{20}$ or 0.05

25% of the students worked 3 hours per day.

Frequency Distribution

- Cumulative relative frequency: the accumulation of the previous relative frequencies
- To find the cumulative relative frequencies, add all the previous relative frequencies to the relative frequency for the current row

DATA VALUE	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
2	3	$\frac{3}{20}$ or 0.15	0.15
3	5	$\frac{5}{20}$ or 0.25	0.15 + 0.25 = 0.40
4	3	$\frac{3}{20}$ or 0.15	0.40 + 0.15 = 0.55
5	6	$\frac{6}{20}$ or 0.30	0.55 + 0.30 = 0.85
6	2	$\frac{2}{20}$ or 0.10	0.85 + 0.10 = 0.95
7	1	$\frac{1}{20}$ or 0.05	0.95 + 0.05 = 1.00

55% of the students worked 4 hours or less per day.

Frequency Distribution

- The *X* values in a frequency distribution table represent the scale of measurement, *not* the actual set of scores
  - E.g., the *X* column lists a value of *X* = 5, but the frequency column indicates that no one actually had a score of *X* = 5

X	*f*
10	2
9	5
8	7
7	3
6	2
5	0
4	1

Frequency Distribution

- If you have a large number of values that cover a wide range, the *ungrouped frequency distribution* is not appropriate.
  - E.g., 1000 observations that range from 0 to 500
- Remember: The purpose for constructing a table is to obtain a relatively simple, organized picture of the data

- In this case, use a *grouped* frequency distribution
  - Grouping the scores into intervals and then listing the intervals in the table instead of listing each individual score

Frequency Distribution

HEIGHTS (INCHES)	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
59.95–61.95	5	$\frac{5}{100} = 0.05$	0.05
61.95–63.95	3	$\frac{3}{100} = 0.03$	0.05 + 0.03 = 0.08
63.95–65.95	15	$\frac{15}{100} = 0.15$	0.08 + 0.15 = 0.23
65.95–67.95	40	$\frac{40}{100} = 0.40$	0.23 + 0.40 = 0.63
67.95–69.95	17	$\frac{17}{100} = 0.17$	0.63 + 0.17 = 0.80
69.95–71.95	12	$\frac{12}{100} = 0.12$	0.80 + 0.12 = 0.92
71.95–73.95	7	$\frac{7}{100} = 0.07$	0.92 + 0.07 = 0.99
73.95–75.95	1	$\frac{1}{100} = 0.01$	0.99 + 0.01 = 1.00
	Total = 100	**Total = 1.00**	

Frequency Distribution

- Guidelines in the construction of a grouped frequency distribution table
  1. Have about 10 class intervals
     - If a table has many more than 10 intervals, it becomes cumbersome and defeats the purpose of a frequency distribution table.
     - If you have too few intervals, you begin to lose information about the distribution of the scores.

Frequency Distribution

- Guidelines in the construction of a grouped frequency distribution table
  2. The width of each interval should be a relatively simple number
     - E.g., 2, 5, 10, 20
     - These numbers are easy to understand and make it possible for someone to see quickly how you have divided the range of scores.

X	*f*
90–94	3
85–89	4
80–84	5
75–79	4
70–74	3
65–69	1
60–64	3
55–59	1
50–54	1

Frequency Distribution

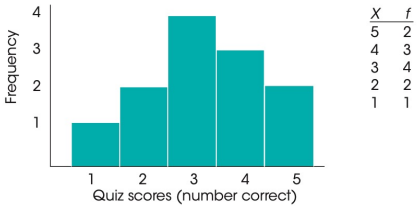
- Guidelines in the construction of a grouped frequency distribution table
  3. All intervals should be the same width
     - Cover the range of scores completely with no gaps and no overlaps
     - Any particular score belongs in exactly one interval

Graphs

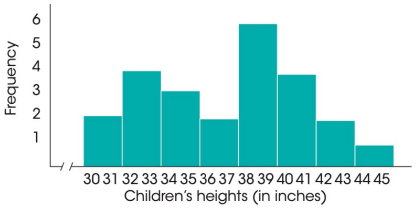
- Sometimes we want to represent a distribution of scores as a graph rather than as a table.
- A frequency distribution graph is basically a picture of the information available in a frequency distribution table.

Histogram

- A graph of a frequency distribution in which a rectangular bar is drawn over each value on the x axis
- Classes are plotted on the x axis
- Frequency is plotted on the y axis



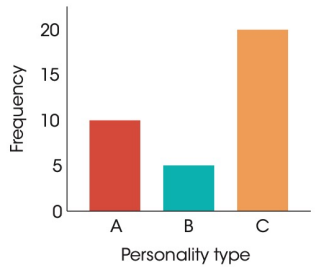
X	f
5	2
4	3
3	4
2	2
1	1



X	f
44–45	1
42–43	2
40–41	4
38–39	6
36–37	2
34–35	3
32–33	4
30–31	2

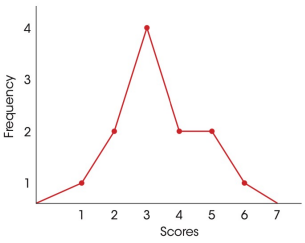
Bar Graph

- A type of histogram used to graph *qualitative* data
- Each bar is separated from its neighbors
  - The space between bars emphasizes that the scale consists of separate, distinct categories

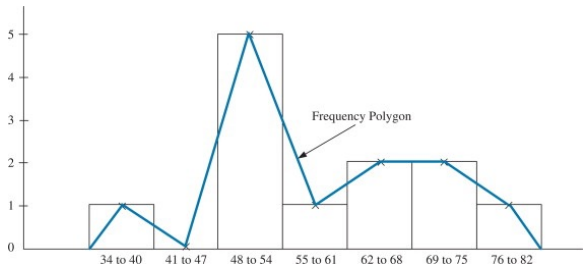


Frequency Polygon

- A line graph of a frequency distribution in which classes are plotted on the x axis and frequencies are plotted on the y axis
- A line graph version of a histogram



X	f
6	1
5	2
4	2
3	4
2	2
1	1



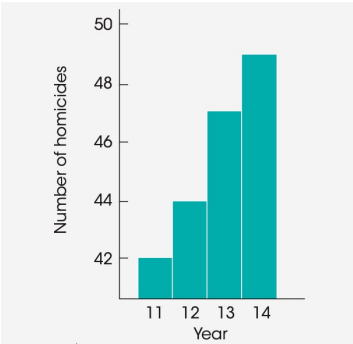
The Use and Misuse Of Graphs

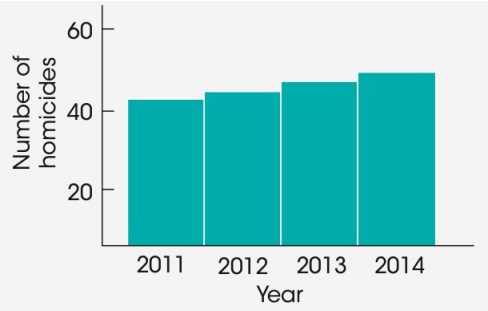
- Although graphs are intended to provide an accurate picture of a set of data, they can be used to exaggerate or misrepresent a set of scores.
- Result from failing to follow the basic rules for graph construction

Year	Number of Homicides
2011	42
2012	44
2013	47
2014	49

The Use and Misuse Of Graphs

- Although graphs are intended to provide an accurate picture of a set of data, they can be used to exaggerate or misrepresent a set of scores.





Learning Check

- For this distribution, how many individuals had scores lower than $X = 20$?

a. 2
b. 3
c. 4
d. cannot be determined

X	f
24–25	2
22–23	4
20–21	6
18–19	3
16–17	1

Learning Check

- The seminar rooms in the library are identified by letters (A, B, C, and so on). A professor records the number of classes held in each room during the fall semester. If these values are presented in a frequency distribution graph, what kind of graph would be appropriate?

a. a histogram
b. a polygon
c. a histogram or a polygon
d. a bar graph

Descriptive Statistics

Measures of the Location of the Data

Hmmmm...

- Individual scores, or *X* values, are called raw scores. By themselves, raw scores do not provide much information.
  - E.g., your score on an exam is *X* = 43
  - How well you did relative to other students?
- It is desirable to transform them into a more meaningful form
  - Median, percentiles and quartiles

Median

- The middle number
- Found by ordering all data points (smallest to largest) and picking out the one in the middle
- 3, 1, 4, 3, 7, 6, 8, 2, 5
- Ordering: 1, 2, 3, 3, 4, 5, 6, 7, 8

Median

- The middle number
- Found by ordering all data points (smallest to largest) and picking out the one in the middle
- 3, 1, 4, 3, 7, 6, 8, 2
- Ordering: 1, 2, 3, 3, 4, 6, 7, 8
- If there are two middle numbers, taking the mean (average) of those two numbers
- (3 + 4) / 2 = 3.5

Median

- Formula for finding the location/index of the median in an ordered data

- $i = \frac{1}{2}(\mathrm{n} + 1)$
- If i is an integer, then the *median* is the data value in the i^{th} position in the ordered set of data.
- If i is not an integer, then round i up and round i down to the nearest integers. Average the two data values in these two positions in the ordered data set to get the *median*.

Percentiles

- Median divides ordered data into two halves
- Percentiles divide ordered data into hundredths
- Median = 50^{th} percentile
- Formula for finding the k^{th} percentile
  - $i = \frac{k}{100}(\mathrm{n} + 1)$
  - If i is an integer, then the k^{th} percentile is the data value in the i^{th} position in the ordered set of data.
  - If i is not an integer, then round i up and round i down to the nearest integers. Average the two data values in these two positions in the ordered data set.

Quartiles

- Median divides ordered data into two halves
- Percentiles divide ordered data into hundredths
- Quartiles divide ordered data into quarters
  - E.g., 1st: 25th percentile, 3rd: 75th percentile
- Median = 50th percentile = 2nd quartile
- Formula for finding the kth quartile
  - $i = \frac{k}{4}(\mathrm{n}+1)$
  - If i is an integer, then the k^{th} quartile is the data value in the i^{th} position in the ordered set of data.
  - If i is not an integer, then round i up and round i down to the nearest integers. Average the two data values in these two positions in the ordered data set.

Quartiles

- Interquartile range (IQR): a number that indicates the spread of the middle half or the middle 50% of the data
- $IQR=Q_3-Q_1$
- Help to determine potential outliers
  - A data point that is significantly different from the other data points
  - Require further investigation: errors or some kind of abnormality or a key to understand the data
- A value is suspected to be a potential outlier if
  - it is less than (1.5)(IQR) below the first quartile
  - or more than (1.5)(IQR) above the third quartile

Example

- Listed are 29 ages for Academy Award winning best actors in order from smallest to largest.
- 18; 21; 22; 25; 26; 27; 29; 30; 31; 33; 36; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 78
- Find the median.
- $i = \frac{1}{2}(\mathrm{n} + 1) = \frac{1}{2}(29 + 1) = 15$
- 15 is an integer, and the data value in the 15th position in the ordered data set is 47.
- The median is 47 years old.

Example

- Listed are 29 ages for Academy Award winning best actors in order from smallest to largest.
- 18; 21; 22; 25; 26; 27; 29; 30; 31; 33; 36; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 78
- Find the 70th percentile.
- $i = \frac{k}{100}(\mathrm{n} + 1) = \frac{70}{100}(29 + 1) = 21$
- 21 is an integer, and the data value in the 21th position in the ordered data set is 64.
- The 70th percentile is 64 years old.

Example

- Listed are 29 ages for Academy Award winning best actors in order from smallest to largest.
- 18; 21; 22; 25; 26; 27; 29; 30; 31; 33; 36; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 78
- Find the 83rd percentile.
- $i = \frac{k}{100}(\mathrm{n} + 1) = \frac{83}{100}(29 + 1) = 24.9$
- 24.9 is NOT an integer.
- Round it down to 24 and up to 25.
- The age in the 24th position is 71 and the age in the 25th position is 72.
- Average 71 and 72 → 71.5
- The 83rd percentile is 71.5 years old.

Example

- Listed are 29 ages for Academy Award winning best actors in order from smallest to largest.
- 18; 21; 22; 25; 26; 27; 29; 30; 31; 33; 36; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 78
- Find the 3rd quartile.
- $i = \frac{k}{4}(\mathrm{n} + 1) = \frac{3}{4}(29 + 1) = 22.5$
- 22.5 is NOT an integer.
- Round it down to 22 and up to 23.
- The age in the 22nd position is 67 and the age in the 23rd position is 69.
- Average 67 and 69 → 68
- The 3rd quartile is 68 years old.

Example

- Listed are 29 ages for Academy Award winning best actors in order from smallest to largest.
- 18; 21; 41; 42; 43; 45; 55; 57; 58; 59; 59; 59; 60; 60; 61; 61; 62; 62; 63; 63; 64; 67; 69; 71; 72; 73; 74; 76; 88
- Calculate the IQR and determine if any ages are potential outliers
- $Q_3 = 68$
- Q_1 index: $i = \frac{k}{4}(\mathrm{n}+1) = \frac{1}{4}(29+1) = 7.5$
- $Q_1 = (55 + 57) / 2 = 56$
- IQR = 68 – 56 = 12

Example

- Listed are 29 ages for Academy Award winning best actors in order from smallest to largest.
- 18; 21; 41; 42; 43; 45; 55; 57; 58; 59; 59; 59; 60; 60; 61; 61; 62; 62; 63; 63; 64; 67; 69; 71; 72; 73; 74; 76; 88
- Calculate the IQR and determine if any ages are potential outliers
- $Q_1 = 56$; $Q_3 = 68$
- IQR = 68 – 56 = 12
- 1.5 * IQR = 18
- $Q_1 – 1.5*IQR = 56 – 18 = 38$; $Q_3 + 1.5*IQR = 68 + 18 = 86$
- 18 and 21 are less than 38; 88 is more than 86

Interpreting Percentiles, Quartiles, and Median

- To score in the 90th percentile of an exam does not mean, necessarily, that you received 90% on a test.
- It means that 90% of test scores are (the same or) less than your score and 10% of the test scores are (the same or) greater than your test score.

Interpreting Percentiles, Quartiles, and Median

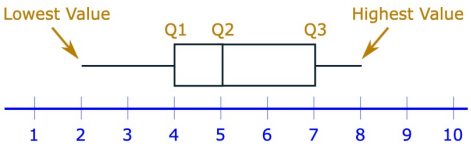
- A percentile may or may not correspond to a value judgment about whether it is “good” or “bad.”
  - The interpretation of whether a certain percentile is “good” or “bad” depends on the context of the situation to which the data applies. In many situations, there is no value judgment that applies.
  - E.g., on a timed math test, the first quartile for time it took to finish the exam was 35 minutes. Interpret the first quartile in the context of this situation.
    - 25% of students finished the exam in 35 minutes or less.
    - 75% of students finished the exam in 35 minutes or more.
    - A low percentile could be considered good, as finishing more quickly on a timed exam is desirable. (If you take too long, you might not be able to finish.)

Interpreting Percentiles, Quartiles, and Median

- A percentile may or may not correspond to a value judgment about whether it is "good" or "bad."
  - The interpretation of whether a certain percentile is "good" or "bad" depends on the context of the situation to which the data applies. In many situations, there is no value judgment that applies.
  - E.g., at a community college, it was found that the 30^{th} percentile of credit units that students are enrolled for is 7 units. Interpret the 30^{th} percentile in the context of this situation.
    - 30% of students are enrolled in 7 or fewer credit units.
    - 70% of students are enrolled in 7 or more credit units.
    - In this example, there is no "good" or "bad" value judgment associated with a higher or lower percentile. Students attend community college for varied reasons and needs, and their course load varies according to their needs.

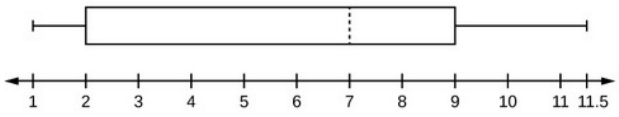
Box Plots

- Box-and-whisker plots or box-whisker plots
- Give a good graphical image of the concentration of the data
- Show how far the extreme values are from most of the data
- A box plot is constructed from five values:
  - The minimum value
  - The first quartile
  - The median
  - The third quartile
  - The maximum value



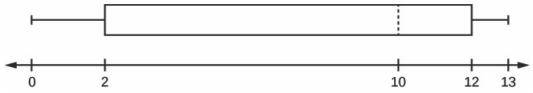
Box Plots

- The middle 50 percent of the data fall inside the box
- The “whiskers” extend from the ends of the box to the smallest and largest data values
- The median or second quartile can be between the first and third quartiles, or it can be one, or the other, or both



Example

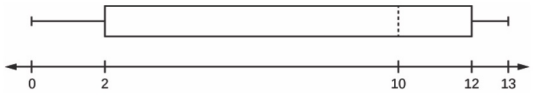
- Given the following box plot:



- Which quarter has the smallest spread of data? What is that spread?
  - The last 25% (Q_3 to max), spread from 12–13

Example

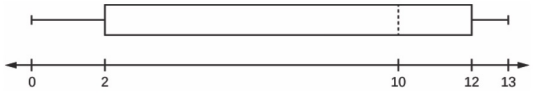
- Given the following box plot:



- Which quarter has the largest spread of data? What is that spread?
  - The second 25% (Q_1 to Q_2), spread from 2–10

Example

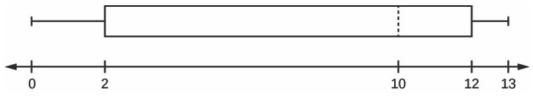
- Given the following box plot:



- Find the interquartile range (*IQR*)
  - $IQR = Q_3 - Q_1 = 12 - 2 = 10$

Example

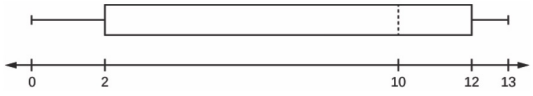
- Given the following box plot:



- Are there more data in the interval 5–10 or in the interval 10–13? How do you know this?
  - There are more data in the interval 10–13. This interval holds 50% of the distribution of data.
  - There is only 25% in the interval 2–10 therefore the interval 5–10 will hold less than 25%.

Example

- Given the following box plot:



- Which interval has the fewest data in it? How do you know this?
  A. 0—2
  B. 2—4
  C. 10—12
  D. 12—13
  E. Need more information

The interval from 2–4 has the fewest proportion of the data as it is less than 25%, it is a subset of the interval 2–10. All of the other intervals listed hold 25% of the data.

Descriptive Statistics

Measures of the Center of the Data

Measures of the Center of the Data

- The general purpose of descriptive statistical methods is to organize and summarize a set of scores
- The most common method for summarizing and describing a distribution is to find a single value that …
  - defines the average score
  - can serve as a typical example to represent the entire distribution
- In statistics, the concept of an average or representative score is called *central tendency*.
  - The goal of central tendency is to find the single score that is most typical or most representative of the entire group.

Mode

- The most frequently occurring observation/measurement
- Both quantitative and qualitative data
  - 1, 1, 2, 3, 3, 3, 3, 3, 3, 4, 4, 5, 5, 5, 6, 8, 9
  - red, red, green, green, green, green, blue, yellow, yellow
- Remember: The MODE is the value, NOT the frequency

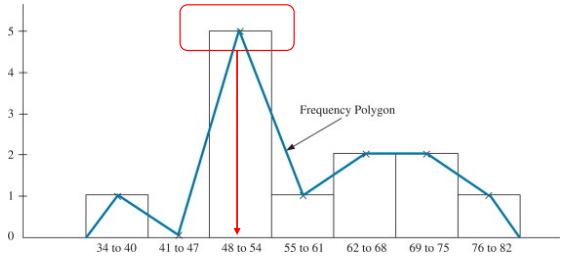
Mode

- The most frequently occurring observation/measurement
- Both quantitative and qualitative data
- In a frequency distribution, look for the highest frequency

DATA VALUE (*X*)	FREQUENCY (*f*)
2	3
3	5
4	3
5	6
6	2
7	1

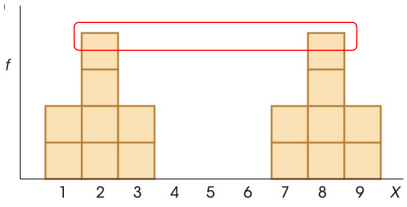
Mode

- The most frequently occurring observation/measurement
- Both quantitative and qualitative data
- In a graph, look for the highest bar in a histogram or highest point in a polygon



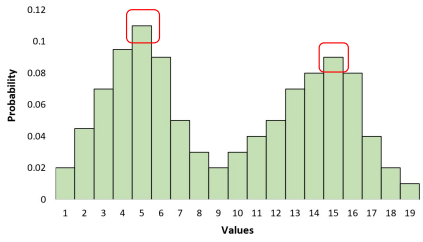
Mode

- The most frequently occurring observation/measurement
- Both quantitative and qualitative data
- Distributions with two peaks are bimodal (have two modes)



Mode

- The most frequently occurring observation/measurement
- Both quantitative and qualitative data
- Distributions with two peaks are bimodal (have two modes)
  - Even if the peaks are not exactly the same height



Median

- The middle value when observations are ordered from least to most
  - Midpoint: divides scores into two equal-sized groups
  - Half of the *scores* lie above the median; half of the *scores* lie below the median
- Formula for finding the location/index of the median
  - $i = \frac{1}{2}(\mathrm{n} + 1)$
  - If i is an integer, then the *median* is the data value in the i^{th} position in the ordered set of data.
  - If i is not an integer, then round i up and round i down to the nearest integers. Average the two data values in these two positions in the ordered data set.
- Remember: The MEDIAN is the value, NOT the location/index

Mean

- The most commonly used and most useful average
- The sum of the scores in a distribution divided by the number of scores
- $Mean = \frac{sum\ of\ all\ observations}{number\ of\ all\ observations} = \frac{\sum X}{n}$
- Observations can be added in any order

Summation Sign ($\sum$ "sigma")

- Σ: 'the sum of"
- $\sum X$: to add all the scores for variable X
- $\sum X^2$: first square each value of X, then add all of those squared values
- $\sum(X + 1)$: first add 1 to each value of X, then add all of those new values
- $\sum(X + 1)^2$?
  - Add 1 to each value of X
  - Square each of those new values
  - Add all of those squared values

Summation Sign (∑ “sigma”)

- X: 4, 6, 0, 3, 2
- Find $\sum X$
- $\sum X = 15$

Summation Sign (∑ “sigma”)

- X: 4, 6, 0, 3, 2
- Find $\sum X^2$
- $\sum X^2 = 65$

Summation Sign (∑ "sigma")

- X: 4, 6, 0, 3, 2
- Find $\sum(X+1)^2$
- $\sum(X+1)^2 = 100$

Mean

- Formula for sample mean
  - $\bar{X} = \frac{\sum X}{n}$
  - In psychology (i.e., APA format), *M* is the symbol for sample mean (not median as shown in the textbook)
  - For this class, both $\bar{X}$ and *M* represent sample mean
- Formula for population mean
  - $\mu = \frac{\sum X}{N}$

Mean

- Find mean from frequency distribution table
- $\bar{X} = \frac{\sum fX}{\sum f}$

x	*f*
3	1
4	2
5	2
6	2
7	4
8	1

Mean

- Find mean from frequency distribution table
- $\bar{X} = \frac{\sum fX}{\sum f}$

x	*f*	*fx*
3	1	3
4	2	8
5	2	10
6	2	12
7	4	28
8	1	8

Mean

- Find mean from frequency distribution table
- $\bar{X} = \frac{\sum fX}{\sum f} = \frac{69}{\sum f}$

X	*f*	*fX*
3	1	3
4	2	8
5	2	10
6	2	12
7	4	28
8	1	8

Mean

- Find mean from frequency distribution table
- $\bar{X} = \frac{\sum fX}{\sum f} = \frac{69}{\sum f} = \frac{69}{12} = 5.75$

X	*f*	*fX*
3	1	3
4	2	8
5	2	10
6	2	12
7	4	28
8	1	8

Mean

- A sample of *n*=7 scores has a mean of *M*=9
- What is $\sum X$?

- $M = \frac{\sum X}{n}$
- $9 = \frac{\sum X}{7}$
- $\sum X = 9 \times 7 = 63$

Mean

- A sample of *n*=8 scores has a mean of *M*=10
- Include one new observation with a score of *X*=1
- What is the new mean?

- Old *n*=8, old *M*=10
- New *n*=9, new *M*= ?
- $M = \frac{\sum X}{n}$
- New $\sum X$ =?

Mean

- A sample of *n*=8 scores has a mean of *M*=10
- Include one new observation with a score of *X*=1
- What is the new mean?

- Old *n*=8, old *M*=10
- New *n*=9, new *M*= ?
- $M = \frac{\sum X}{n}$
- $\text{New} \sum X = \text{Old} \sum X + \text{NewValue} = 8 \times 10 + 1 = 81$

Mean

- A sample of *n*=8 scores has a mean of *M*=10
- Include one new observation with a score of *X*=1
- What is the new mean?

- Old *n*=8, old *M*=10
- New *n*=9, new *M*= ?
- $New \sum X = Old \sum X + NewValue = 8 \times 10 + 1 = 81$
- $New\ M = \frac{New \sum X}{New\ n} = \frac{81}{9} = 9$

Mean

- Open-ended distribution
- It is impossible to compute a mean for these data because you cannot find $\sum X$
- However, you can find the median

Number of Pizzas (*X*)	*f*
5 or more	3
4	2
3	2
2	3
1	6
0	4

Report Measures of Central Tendency

- *Publication Manual of the American Psychological Association* (2020)
- APA style
- The APA style uses the letter *M* as the symbol for the sample mean.
- Thus, a study might state:
  - The treatment group showed fewer errors (*M* = 2.56) on the task than the control group (*M* = 11.76).

Report Measures of Central Tendency

- *Publication Manual of the American Psychological Association* (2020)
- APA style
- When there are many means to report, tables with headings provide an organized and more easily understood presentation

The mean number of errors made on the task for treatment and control groups according to gender.

	Treatment	Control
Females	1.45	8.36
Males	3.83	14.77

Report Measures of Central Tendency

- *Publication Manual of the American Psychological Association* (2020)
- APA style
- The median can be reported using the abbreviation *Mdn,* as in "Mdn = 8.5 errors"
- Or it can simply be reported in narrative text, as follows:
  - The median number of errors for the treatment group was 8.5, compared to a median of 13 for the control group.

Report Measures of Central Tendency

- *Publication Manual of the American Psychological Association* (2020)
- APA style
- No special symbol or convention for reporting the mode
- If mentioned at all, the mode is usually just reported in narrative text

Descriptive Statistics

Skewness and Central Tendency

Open-ended Distribution

- It is impossible to compute a mean for these data because you cannot find $\sum X$
- However, you can find the median ?
  - Midpoint (location/index)
  - 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 2, 2, 2, 3, 3, 4, 4
  - 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 2, 2, 2, 3, 3, 4, 4, 5, 5, 5
  - 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 2, 2, 2, 3, 3, 4, 4, 5, 6, 8
  - 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 2, 2, 2, 3, 3, 4, 4, 5, 60, 80
  - 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 2, 2, 2, 3, 3, 4, 4, 5, 600, 800
  - …

Number of Pizzas (*X*)	*f*
5 or more	3
4	2
3	2
2	3
1	6
0	4

Open-ended Distribution

- It is impossible to compute a mean for these data because you cannot find $\sum X$
- However, you can find the median ?
  - Midpoint (location/index)
  - 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 2, 2, 2, 3, 3, 4, 4, 5, 5, 5
  - $i = \frac{1}{2}(\mathrm{n}+1) = \frac{1}{2}(\sum f + 1) = \frac{1}{2}(20+1) = 10.5$
  - $Mdn = \frac{(1+2)}{2} = 1.5$

Number of Pizzas (*X*)	*f*
5 or more	3
4	2
3	2
2	3
1	6
0	4

Mean

- X: 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 2, 2, 2, 3, 3, 4, 4, 5, 5, 5
- Find the mean
- $M = \frac{\sum X}{n} = \frac{41}{20} = 2.05$

Mean

- X: 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 2, 2, 2, 3, 3, 4, 4, 5, 5, 9
- Find the mean
- $M = \frac{\sum X}{n} = \frac{45}{20} = 2.25$

Mean

- X: 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 2, 2, 2, 3, 3, 4, 4, 5, 5, 15
- Find the mean
- $M = \frac{\sum X}{n} = \frac{51}{20} = 2.55$

Mean and Median

- Mean: based on all of the data
  - Adding or removing even one score usually changes the mean
- Median: not influenced by very deviant scores
  - The median is generally a better measure of the center when there are extreme values or outliers because it is not affected by the precise numerical values of the outliers

Mean and Median

Geography	Canada[4] (map)				
Age group	16 years and over				
Sex	Both sexes				
Income source	Total income				
Statistics	**2015**	**2016**	**2017**	**2018**	**2019**
	Number				
Number of persons (x 1,000)	29,052[A]	29,341[A]	29,711[A]	30,046[A]	30,631[A]
Number with income (x 1,000)	28,359[A]	28,601[A]	29,035[A]	29,510[A]	29,984[A]
	2019 constant dollars				
Aggregate income (x 1,000,000)	1,348,210[A]	1,353,085[A]	1,414,478[A]	1,444,435[A]	1,469,998[A]
Average income (excluding zeros)	47,500[A]	47,300[A]	48,700[A]	48,900[A]	49,000[A]
Median income (excluding zeros)	35,300[A]	35,600[A]	36,500[A]	37,100[A]	37,800[A]

Mean

- X: 0, 2, 3, 4, 6

- Find the mean
- $M = \frac{\sum X}{n} = \frac{15}{5} = 3$

- Find $\sum(X - M)$
- $\sum(X - M) = (0 - 3) + (2 - 3) + (3 - 3) + (4 - 3) + (6 - 3)$
- $\sum(X - M) = -3 - 1 + 0 + 1 + 3 = 0$
- Mean as a balancing point

Mean

- X: 0, 2, 3, 4, 6

- Find $\sum(X-M)^2$
- $\sum(X-M)^2 = (0-3)^2+(2-3)^2+(3-3)^2+(4-3)^2+(6-3)^2$
- $\sum(X-M)^2 = 9+1+0+1+9 = 20$

- Find $(\sum X)^2$ vs. $\sum X^2$

Mean

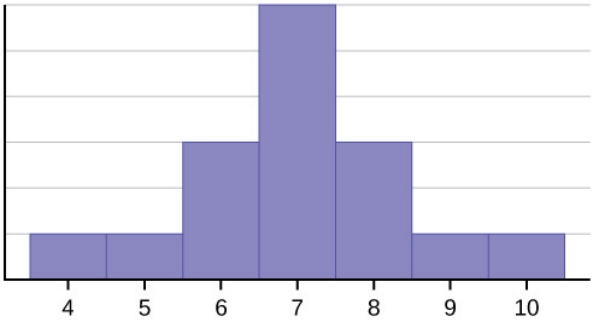
- X: 0, 2, 3, 4, 6

- Find $(\sum X)^2$ vs. $\sum X^2$
- $(\sum X)^2 = (0+2+3+4+6)^2 = 15^2 = 225$
- $\sum X^2 = 0^2+2^2+3^2+4^2+6^2 = 65$

Central Tendency and the Shape of the Distribution

- Symmetrical distribution
  - The right-hand side of the graph is a mirror image of the left-hand side

- Mode = ?
- Median = ?
- Mean = ?

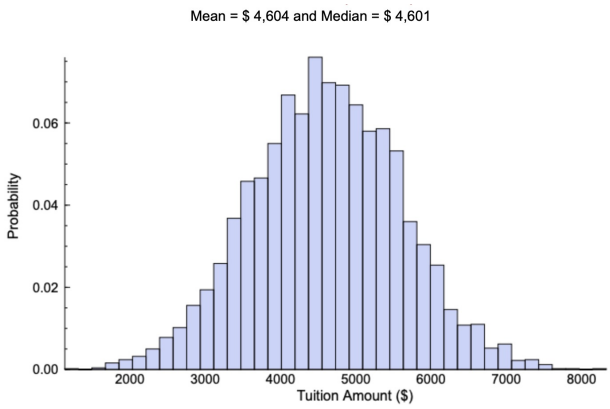


Central Tendency and the Shape of the Distribution

- Symmetrical distribution
  - The right-hand side of the graph is a mirror image of the left-hand side
  - For a perfectly (unimodal) symmetrical distribution, the mean, the median and the mode are the same
  - If a distribution is roughly symmetrical, but not perfect, the mean and median will be close together in the center of the distribution

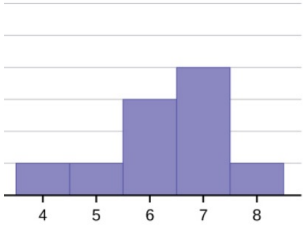
Central Tendency and the Shape of the Distribution

- Roughly symmetrical distribution



Central Tendency and the Shape of the Distribution

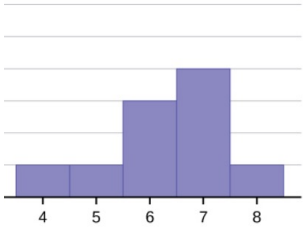
- Skewed distributions
  - The right-hand side seems “chopped off” compared to the left side.
  - A distribution of this type is called skewed to the left (negatively skewed) because it is pulled out to the left.



- X: 4; 5; 6; 6; 6; 7; 7; 7; 7; 8
- Mode=?
- Median=?
- Mean=?

Central Tendency and the Shape of the Distribution

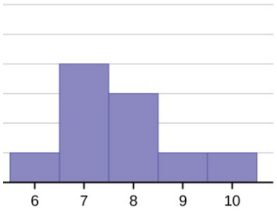
- Skewed distributions
  - The right-hand side seems “chopped off” compared to the left side.
  - A distribution of this type is called skewed to the left (negatively skewed) because it is pulled out to the left.



- X: 4; 5; 6; 6; 6; 7; 7; 7; 7; 8
- Mode=7
- Median=6.5
- Mean=6.3
- The mean and the median both reflect the skewing, but the mean reflects it more so.

Central Tendency and the Shape of the Distribution

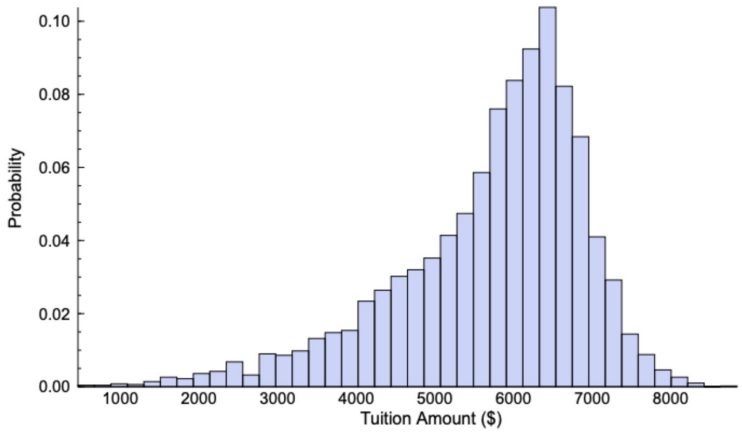
- Skewed distributions
  - Skewed to the right (positively skewed)



- X: 6; 7; 7; 7; 7; 8; 8; 8; 9; 10
- Mode=7
- Median=7.5
- Mean=7.7
- The mean and the median both reflect the skewing, but the mean reflects it more so.

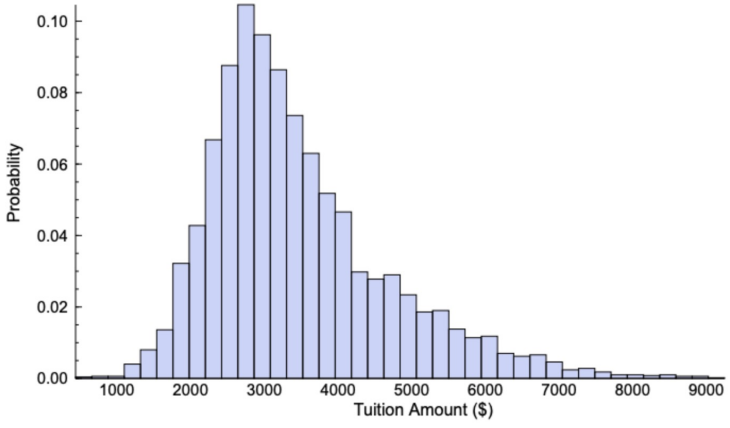
Central Tendency and the Shape of the Distribution

Mean = $ 5,720 and Median = $ 5,994

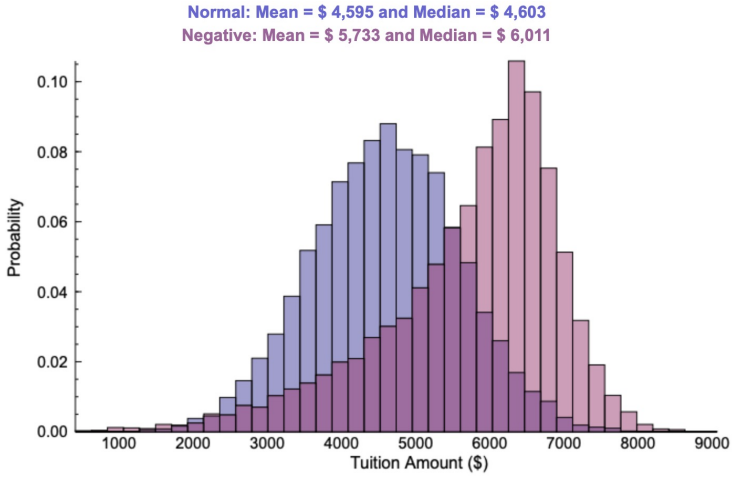


Central Tendency and the Shape of the Distribution

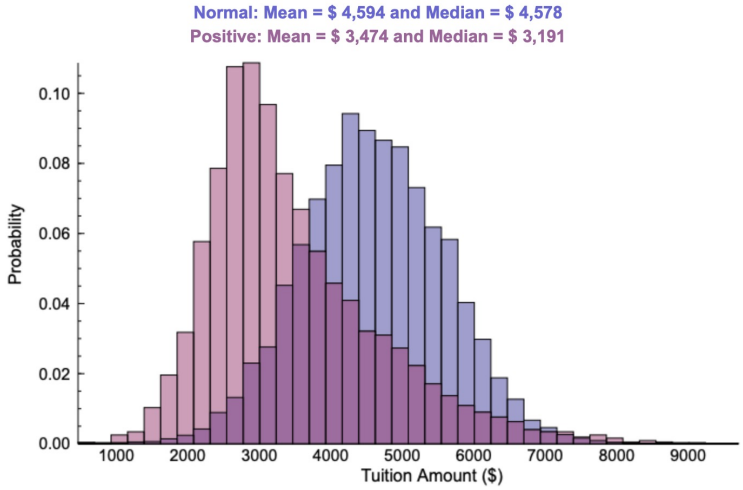
Mean = $ 3,461 and Median = $ 3,182



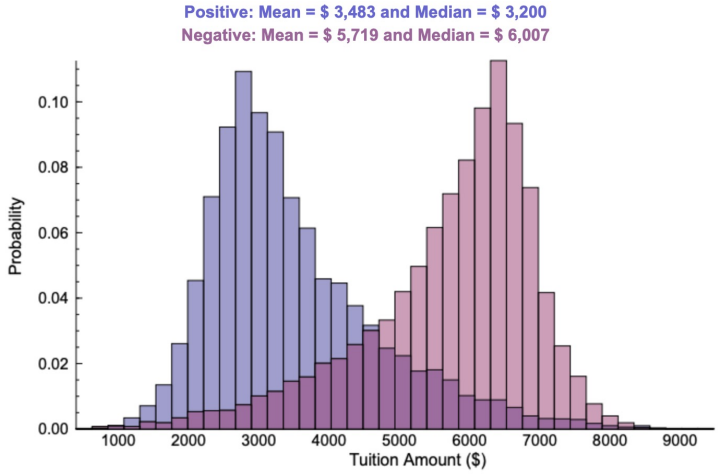
Central Tendency and the Shape of the Distribution



Central Tendency and the Shape of the Distribution

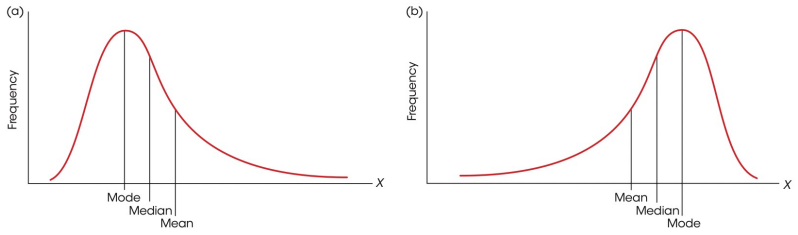


Central Tendency and the Shape of the Distribution



Central Tendency and the Shape of the Distribution

- Skewed distributions
  - Generally if the distribution of data is skewed to the left, the mean is less than the median, which is often less than the mode.
  - If the distribution of data is skewed to the right, the mode is often less than the median, which is less than the mean.



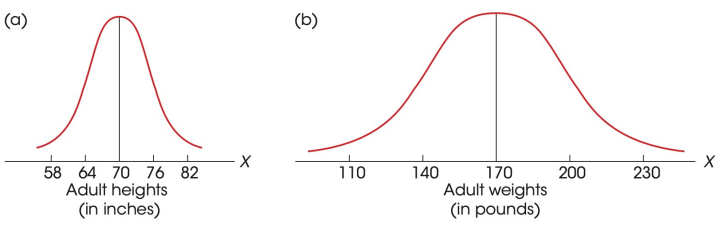
Descriptive Statistics

Measures of Variability

Measures of Variability

- We have discussed *measures of central tendency*
- Now we need measures to describe the degree to which observations:
  - Cluster together

  OR
  - Differ or deviate



Measures of Variability

- Measures that indicate the degree of difference among observations

Range

- The distance from the lowest score to the highest score
- $Range = X_{max} - X_{min}$

Range

- Listed are 29 ages for Academy Award winning best actors in order from smallest to largest.
- 18; 21; 22; 25; 26; 27; 29; 30; 31; 33; 36; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 78
- What is the range?
- $Range = X_{max} - X_{min} = 78 - 18 = 60$

Range

- The distance from the lowest score to the highest score
- $Range = X_{max} - X_{min}$
- Problems with the range
  - Based on only two observations
  - As number of observations increases, range increases
    - Why?
    - There is an increased probability of extreme values

Interquartile Range (IQR)

- Range for the middle 50% of the observations
  - Chop off top 25% of observations
  - Chop off bottom 25% of observations
- $IQR = Q_3 - Q_1$
- Advantage of IQR over range
  - Not sensitive to extreme values
- Problems with the IQR
  - Based on only two values

Deviation

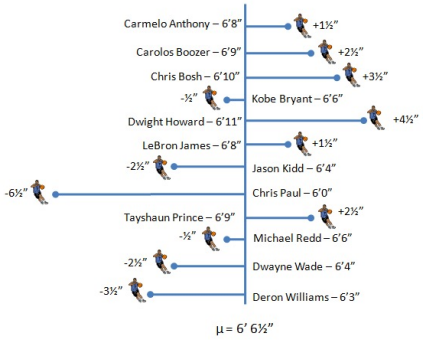
- We could also describe variability by describing how much *each* score differs or deviates from the mean
  - $X - \mu$
- If we want to describe how observations deviate from the mean, why not find the average of the deviations?
  - $\frac{\sum(X-\mu)}{N}$?

Average or Mean Deviation

- The heights of the 2008 US Men's Olympic Basketball team and each player's corresponding difference from the mean

- $\frac{1.5+2.5+3.5-0.5+4.5+1.5-2.5-6.5+2.5-0.5-2.5-3.5}{12}$
- $= \frac{0}{12} = 0$

- The mean is the balancing point!

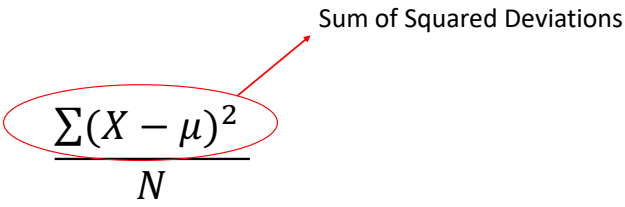


Average or Mean Deviation

- *Average* deviations always equal zero
  - *Average* of deviations tells us nothing
- Solution?
- Get rid of negative signs:

SQUARE EACH DEVIATION

Average Squared Deviation



Sum of Squared Deviations

- Definition formula:

$$SS = \sum (X - \mu)^2$$

Sum of Squared Deviations

- Computation formula:
- Definition formula = Computation formula

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

Sum of Squared Deviations

- X: 0, 2, 3, 4, 6

- Find the *SS* using the definition formula
- $\mu = \frac{0+2+3+4+6}{5} = 3$
- $SS = (0-3)^2+(2-3)^2+(3-3)^2+(4-3)^2+(6-3)^2$
- $SS = 9+1+0+1+9 = 20$

Sum of Squared Deviations

- X: 0, 2, 3, 4, 6

- Find the *SS* using the computational formula
- $SS = \sum X^2 - \frac{(\sum X)^2}{N}$
- $SS = 0^2 + 2^2 + 3^2 + 4^2 + 6^2 - \frac{(0+2+3+4+6)^2}{5}$
- $SS = 0 + 4 + 9 + 16 + 36 - 45 = 20$

Sum of Squared Deviations

- Definition formula vs. Computational formula
- Although the definitional formula is the most direct method for computing *SS,* it can be awkward to use.
  - When the mean is not a whole number, the deviations all contain decimals or fractions, and the calculations become difficult
  - Calculations with decimal values introduce the opportunity for rounding error

Variance

- The mean of the squared deviations
  - The average squared distance from the mean

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

Variance

- Variance reflects the degree of deviation of observations from the mean
- But variance is expressed in SQUARED units
- How do we get back to the original units we started with?
- Take the *square root* of the variance

Standard Deviation

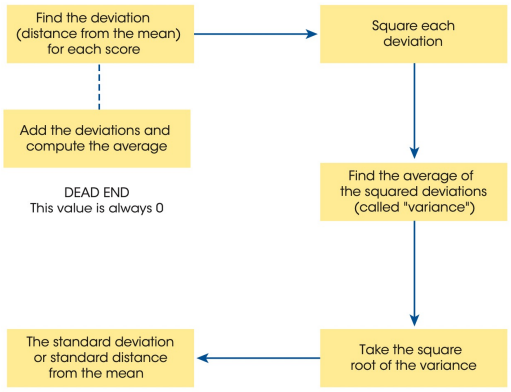
- A measure of the average deviation of the observations from the mean
- The SQUARE ROOT of the variance

- $population\ standard\ deviation = \sigma = \sqrt{\sigma^2} = \sqrt{\frac{SS}{N}}$

Standard Deviation

- If the variance is 2009, what is the standard deviation?

- $\sqrt{2009} \approx 44.82$



Sample Statistics

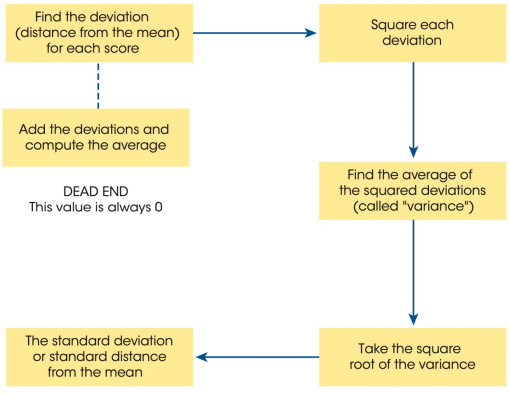
- Sum of Squared Deviations
  - $SS = \sum(X - M)^2$
  - $SS = \sum X^2 - \frac{(\sum X)^2}{n}$
- Variance
  - $s^2 = \frac{SS}{n-1}$

Why n-1 ?

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

Descriptive Statistics

Measures of Variability



Sample Statistics

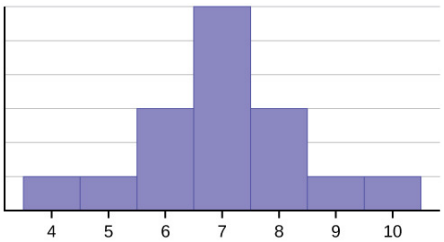
- Sum of Squared Deviations
  - $SS = \sum(X - M)^2$
  - $SS = \sum X^2 - \frac{(\sum X)^2}{n}$
- Variance
  - $s^2 = \frac{SS}{n-1}$

Why n-1 ?

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

Example

- Population
  - X: 4, 5, 6, 6, 6, 7, 7, 7, 7, 7, 7, 8, 8, 8, 9, 10
- Find population standard deviation σ ?



Example

- Population
  - X: 4, 5, 6, 6, 6, 7, 7, 7, 7, 7, 7, 8, 8, 8, 9, 10

- Find population standard deviation σ ?
- $\sigma = \sqrt{\sigma^2} = \sqrt{\frac{SS}{N}}$
- $SS = \sum(X - \mu)^2 = (4-7)^2 + (5-7)^2 + (6-7)^2 + \cdots + (10-7)^2$
- $SS = 9 + 4 + 1 + 1 + 1 + 0 + \cdots + 0 + 1 + 1 + 1 + 4 + 9 = 32$
- $\sigma = \sqrt{\frac{32}{16}} = \sqrt{2} \approx 1.41$

Example

- Population
  - X: 4, 5, 6, 6, 6, 7, 7, 7, 7, 7, 7, 8, 8, 8, 9, 10

- Sample from population
  - X: 6, 6, 7, 7, 7, 7, 8, 8
- Find sample standard deviation s ?
- What if not $\mathrm{n} - 1$?

Example

- Sample from population
  - X: 5, 6, 6, 7, 7, 7, 8, 8, 9
- Find sample standard deviation s ?
- What if not $\mathrm{n}-1$?
- $\mathrm{s}=\sqrt{\frac{SS}{n}}$
- $SS=\sum(X-M)^2\ =(5-7)^2+(6-7)^2+\cdots+(9-7)^2=12$
- $\mathrm{s}=\sqrt{\frac{12}{9}}=\sqrt{1.333\ldots}\approx 1.15$
- $\mathrm{s}<\sigma$

Example

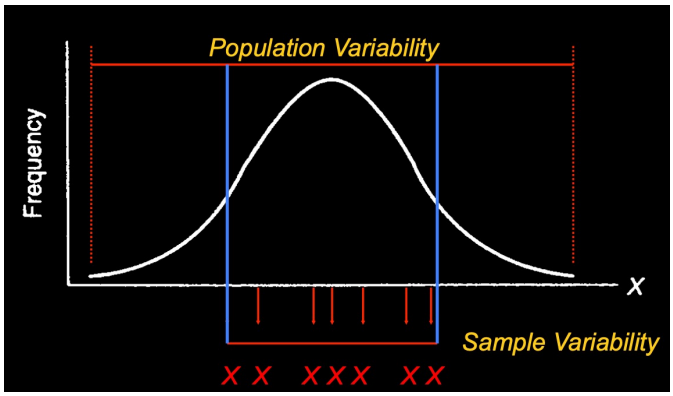
- Sample from population
  - X: 5, 6, 6, 7, 7, 7, 8, 8, 9
- Find sample standard deviation s ?
- $\mathrm{n}-1$
- $\mathrm{s}=\sqrt{\frac{SS}{n-1}}$
- $\mathrm{s}=\sqrt{\frac{12}{9-1}}=\sqrt{1.5}\approx 1.22$
- Closer to σ

Sample Standard Deviation

- Why $n - 1$?

- The sample is not a perfect picture of the population
- In particular, sample variability ***underestimates*** population variability
  - Tend to be smaller than population variance and standard deviation
  - Sample variance and standard deviation are *biased* (i.e., they don't provide an accurate picture of the population)
  - Why?

Sample Standard Deviation



Sample Standard Deviation

- Why $n - 1$?

- "-1" corrects for the bias in the sample variance and standard deviation
- Subtracting 1 *increases* the variance and standard deviation to correct for the *under*estimation

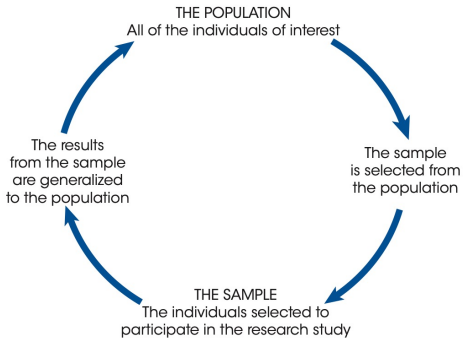
Review

Introduction to Statistics

- **Statistics**: A set of mathematical procedures for *organizing*, *summarizing*, and *interpreting* information
- **Population**: The set of *all* the individuals of interest in a particular study
- **Sample**: A set of individuals selected from a population, usually intended to represent the population in a research study
- **Variable**: A characteristic or condition that *changes* or has *different values* for different individuals
- **Data** (plural): measurements or observations
- **Parameter**: a value, usually a numerical value, that describes a population
- **Statistic**: a value, usually a numerical value, that describes a sample
- **Descriptive statistics**: statistical procedures used to summarize, organize, and simplify data
- **Inferential statistics**: techniques that allow us to study samples and then make *generalizations* about the populations from which they were selected

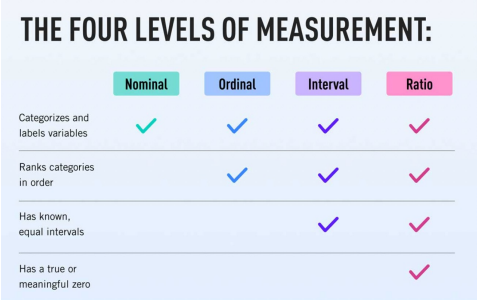
Introduction to Statistics

- Population vs. sample



Introduction to Statistics

- Data types
  - Qualitative vs. quantitative data
  - Quantitative data: result of **counting** (discrete) or **measuring** (continuous)
- Levels/Scales of measurement



Introduction to Statistics

- Correlational vs experimental method
  - Correlational: can demonstrate the existence of a relationship between two variables, but NOT provide an explanation for the relationship
  - Experimental: to demonstrate a cause-and-effect relationship between two variables
- Experimental method
  - Manipulation and control
  - Independent/explanatory vs. dependent/response variable
  - Control lurking variables
    - **Random assignment**: each participant has an equal chance of being assigned to each of the treatment conditions

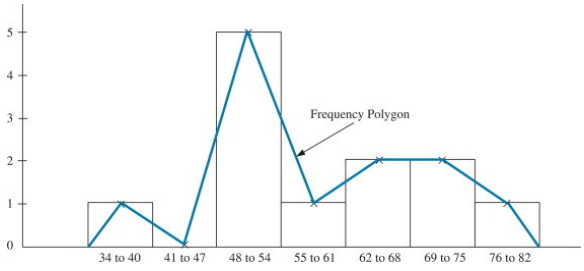
Descriptive Statistics

- Frequency distribution table
  - Sorted collection of observations showing the number of times (or **frequency**) each observation occurs
  - **Relative frequency**: the frequency for an observed value of the data divided by the total number of data values in the sample (proportion/percentage)
  - **Cumulative relative frequency**: the accumulation of the previous relative frequencies

DATA VALUE	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
2	3	$\frac{3}{20}$ or 0.15	0.15
3	5	$\frac{5}{20}$ or 0.25	0.15 + 0.25 = 0.40
4	3	$\frac{3}{20}$ or 0.15	0.40 + 0.15 = 0.55
5	6	$\frac{6}{20}$ or 0.30	0.55 + 0.30 = 0.85
6	2	$\frac{2}{20}$ or 0.10	0.85 + 0.10 = 0.95
7	1	$\frac{1}{20}$ or 0.05	0.95 + 0.05 = 1.00

Descriptive Statistics

- Frequency distribution graphs
  - Histogram
  - Bar graph (qualitative data)
  - Frequency polygon



Descriptive Statistics

- Location of the data

- Median
  - Divides ordered data into two halves
  - $i = \frac{1}{2}(\mathrm{n}+1)$
- Percentiles
  - Divide ordered data into hundredths
  - $i = \frac{k}{100}(\mathrm{n}+1)$
- Quartiles
  - Divide ordered data into quarters
  - $i = \frac{k}{4}(\mathrm{n}+1)$
- Median = 50th percentile = 2nd quartile

Descriptive Statistics

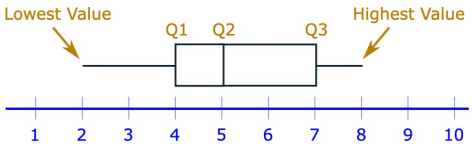
- Location of the data

- Interquartile range (IQR): a number that indicates the spread of the middle half or the middle 50% of the data
- $IQR = Q_3 - Q_1$
- Help to determine potential outliers
  - A value is suspected to be a potential outlier if
    - it is less than (1.5)(IQR) below the first quartile
    - or more than (1.5)(IQR) above the third quartile

Descriptive Statistics

- Location of the data

- Box plot
  - Constructed from five values:
    - The minimum value
    - The first quartile
    - The median
    - The third quartile
    - The maximum value



Descriptive Statistics

- Center of the data

- Central tendency: an average or representative score
- Mode: the most frequently occurring observation/measurement
  - Remember: The MODE is the value, NOT the frequency
  - Bimodal (have two modes)
- Median: midpoint

  Remember: The MEDIAN is the value, NOT the location/index

Descriptive Statistics

- Center of the data

- Mean: the sum of the scores in a distribution divided by the number of scores
- Summation Sign (∑ "sigma")
- Find mean from frequency distribution table
  - $M = \frac{\sum fX}{\sum f} = \frac{69}{\sum f} = \frac{69}{12} = 5.75$

X	f	fX
3	1	3
4	2	8
5	2	10
6	2	12
7	4	28
8	1	8

Descriptive Statistics

- Center of the data

- Mean: the sum of the scores in a distribution divided by the number of scores
- Open-ended distribution
  - Cannot compute the mean
  - May find the median

Number of Pizzas (*X*)	*f*
5 or more	3
4	2
3	2
2	3
1	6
0	4

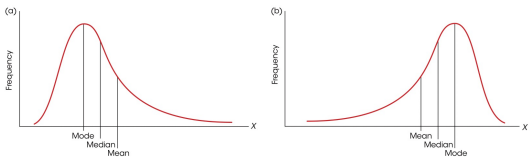
Descriptive Statistics

- Center of the data

- Mean and median
  - Mean: based on all of the data
    - Adding or removing even one score usually changes the mean
  - Median: not influenced by very deviant scores
    - The median is generally a better measure of the center when there are extreme values or outliers because it is not affected by the precise numerical values of the outliers

Descriptive Statistics

- Shape of the distribution
  - Symmetrical
    - For a perfectly (unimodal) symmetrical distribution, the mean, the median and the mode are the same
  - Skewed
    - Generally if the distribution of data is skewed to the left, the mean is less than the median, which is often less than the mode.
    - If the distribution of data is skewed to the right, the mode is often less than the median, which is less than the mean.



Descriptive Statistics

- Measures of variability
  - Measures that indicate the degree of difference among observations

- Range
- Interquartile Range (IQR)
- Deviation: how much *each* score differs or deviates from the mean
  - Population vs sample variance and standard deviation