COMP 7990

Principles and Practices of data analytics

Semester 1, 2022-2023

Dr. Yucheng Jin

Department of Computer Science

Hong Kong Baptist University

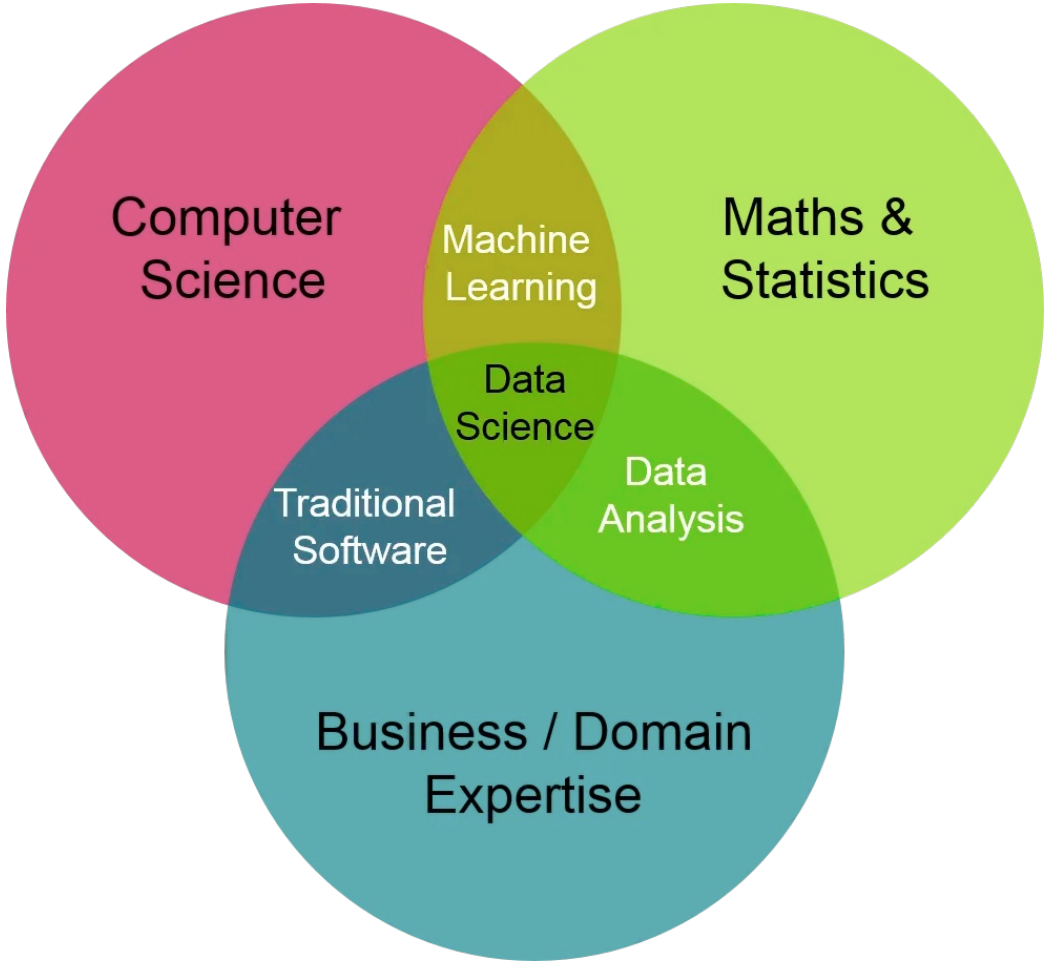
Acknowledge Dr. Lu Zhang for preparing the slides

COMP 7990

Principles and Practices of data analytics

Chapter 1：Statistics

Why Statistics?



Outline

- ❑ Introduction to data
- ❑ Collecting Data: Sampling
- ❑ Descriptive Statistics

Outline

- ❑ Introduction to data
  - ❑ Applications
  - ❑ Concepts
- ❑ Collecting Data: Sampling
- ❑ Descriptive Statistics

Data and statistics

- Statistics is all about DATA
  - Collecting DATA
  - Describing different DATA types - summarizing, visualizing
    - Integer, Decimal, Text
  - Analyzing DATA
- Data are everywhere!
  - The median property price in Hong Kong is now 19.4 times the median income.
  - A single person monthly costs: 7,773.29HKD without rent.
- Statistics is in our daily life
  - health science, weather forecast, stock market, insurance, political campaign...

Scenario 1—Quality Assurance

- Companies make thousands of products every day and each company must make sure that a good quality item is sold.

- However, a company cannot test each and every item that they ship.

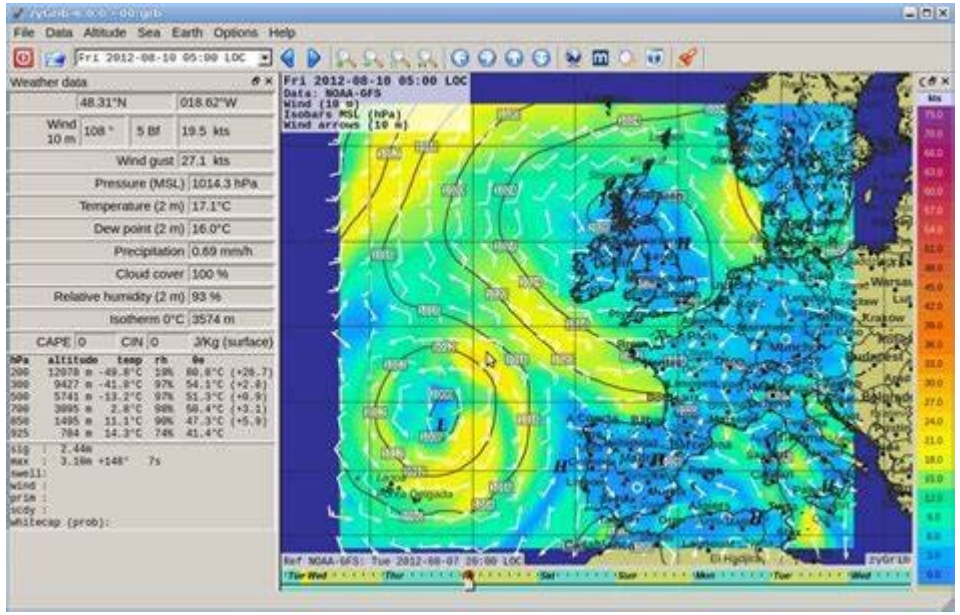
- Thus, a company uses **statistics** to test just a few, called sample, of what they make.
  - If the sample passes quality tests, then the company assumes that all items made in the group are good

Scenario 2— Consumer Goods

- Retailers keep track of everything they sell and use statistics to calculate what to ship, to each store, or warehouse, and when.

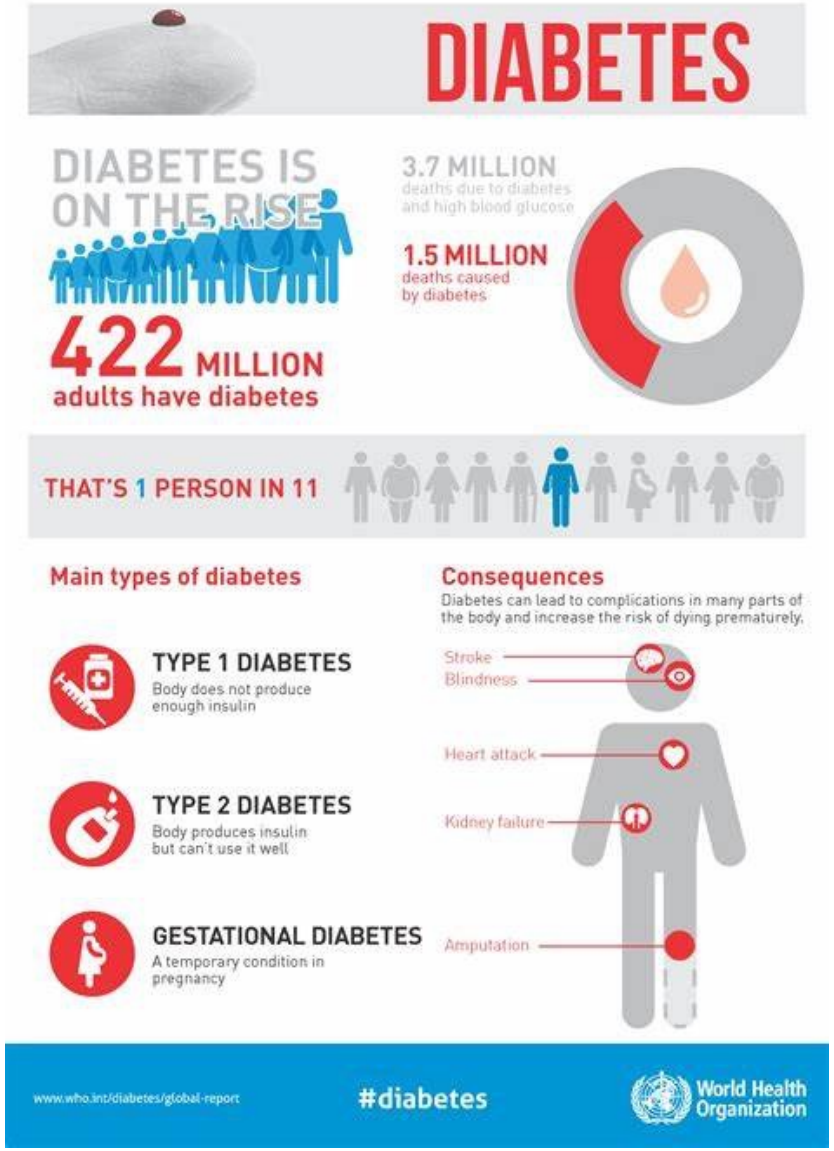
- From analyzing their vast store of information, WalMart decided that people buy strawberry Pop Tarts when a hurricane is predicted in Florida.

- So, they increase their shipment of this product to Florida stores based upon the weather forecast.



Scenario 3 –Disease Prediction

- Often, statistics about disease are reported via various media.
- If the report simply shows the **number of people** who have the disease or who have died from it, it's an interesting fact but it might not mean much to your life.
- When statistics become involved, you have a better idea of how that disease may affect you.
- For example, studies have shown that **85 to 95 percent of lung cancers are smoking related**.
- The statistic should tell you that almost all lung cancers are related to smoking and that if you want to have a good chance of avoiding lung cancer, you shouldn't smoke.



Scenario 4 -- Healthcare

- The steps you take, the flights of stairs you climb and other activities are being tracked by fitness trackers.
- The data gathered through all of this allows you to make better decisions to improve your health.

Class Activity – other applications

Apply University, Medical Studies, Weather forecasts, Political Campaigns, Great jobs……

Outline

- ❑ Introduction to data
  - ❑ Applications
  - ❑ Concepts
- ❑ Collecting Data: Sampling
- ❑ Descriptive Statistics

About Data

- Data are a set of measurements taken on a set of individual units
- Usually data is stored and presented in a dataset, comprised of variables measured on cases
- Example: Below is a first-day survey for students in a statistics class at one university

ID	Gender	Smoke	Award	Exercise	TV	GPA	Pulse	Birth
1	M	No	Olympic	10	1	3.13	54	4
2	F	Yes	Academy	4	7	2.5	66	2
3	M	No	Nobel	14	5	2.55	130	1
4	M	No	Nobel	3	1	3.1	78	1
5	F	No	Nobel	3	3	2.7	40	1
6	F	No	Nobel	5	4	3.2	80	2
7	F	No	Olympic	10	10	2.77	94	1
8	M	No	Olympic	13	8	3.3	77	1
9	F	No	Nobel	3	6	2.8	60	2
10	F	No	Nobel	12	1	3.7	94	8

Cases and Variables -- concept

- We obtain information about cases in a dataset, and generally record the information for each case in a row of a data table.
- A variable is any characteristic that is recorded for each case. The variables generally correspond to the columns in a data table.

A Variable

ID	Gender	Smoke	Award	Exercise	TV	GPA	Pulse	Birth
1	M	No	Olympic	10	1	3.13	54	4
2	F	Yes	Academy	4	7	2.5	66	2
3	M	No	Nobel	14	5	2.55	130	1
A Case 4	M	No	Nobel	3	1	3.1	78	1
5	F	No	Nobel	3	3	2.7	40	1
6	F	No	Nobel	5	4	3.2	80	2
7	F	No	Olympic	10	10	2.77	94	1
8	M	No	Olympic	13	8	3.3	77	1
9	F	No	Nobel	3	6	2.8	60	2
10	F	No	Nobel	12	1	3.7	94	8

Cases and Variables – Class Activity

- Questions
  - How many cases in this dataset?
  - How many variables in this dataset?

ID	Gender	Smoke	Award	Exercise	TV	GPA	Pulse	Birth
1	M	No	Olympic	10	1	3.13	54	4
2	F	Yes	Academy	4	7	2.5	66	2
3	M	No	Nobel	14	5	2.55	130	1
4	M	No	Nobel	3	1	3.1	78	1
5	F	No	Nobel	3	3	2.7	40	1
6	F	No	Nobel	5	4	3.2	80	2
7	F	No	Olympic	10	10	2.77	94	1
8	M	No	Olympic	13	8	3.3	77	1
9	F	No	Nobel	3	6	2.8	60	2
10	F	No	Nobel	12	1	3.7	94	8

In any dataset, it is important to understand exactly what each variable is measuring and how the values are coded.

Cases and Variables – Variables types

- For the data in the table, the first column is ID, to provide an identifier for each of the individuals in the study.
  - Gender: M for male and F for female
  - Smoke: Does the student smoke: yes or no
  - Award: Award the student prefers to win: Academy Award, Olympic gold medal, or Nobel Prize
  - Exercise: Number of hours spent exercising per week
  - TV: Number of hours spent watching television per week
  - GPA: Current grade point average on a 4-point scale
  - Pulse: Pulse rate in number of beats per minute at the time of the survey
  - Birth: Birth order: 1 for first/oldest, 2 for second born, etc.

Cases and Variables – Class Activity

Explain what each variable tells us about the student with **ID 2** in the second row of the table

ID	Gender	Smoke	Award	Exercise	TV	GPA	Pulse	Birth
1	M	No	Olympic	10	1	3.13	54	4
2	F	Yes	Academy	4	7	2.5	66	2
3	M	No	Nobel	14	5	2.55	130	1
4	M	No	Nobel	3	1	3.1	78	1
5	F	No	Nobel	3	3	2.7	40	1
6	F	No	Nobel	5	4	3.2	80	2
7	F	No	Olympic	10	10	2.77	94	1
8	M	No	Olympic	13	8	3.3	77	1
9	F	No	Nobel	3	6	2.8	60	2
10	F	No	Nobel	12	1	3.7	94	8

Example

- Student 1 is a male who does not smoke and who would prefer to win an Olympic gold medal over an Academy Award or a Nobel Prize.
- He says that he exercises 10 hours a week, watches television one hour a week, and that his grade point average is 3.13.
- His pulse rate was 54 beats per minute at the time of the survey, and he is the fourth oldest child in his family.

Exercise

- Student 2 is a _____ who smoke and who would prefer to win _____over _____or _____.
- She says that ____ exercises ____ hours a week, watches television _____ hour a week, and that ____ grade point average is _____.
- Her pulse rate was ____ beats per minute at the time of the survey, and is the _____child in her family.

Cases and Variables – Class Activity

Explain what each variable tells us about the student with ID 2 in the second row of the table

ID	Gender	Smoke	Award	Exercise	TV	GPA	Pulse	Birth
1	M	No	Olympic	10	1	3.13	54	4
2	F	Yes	Academy	4	7	2.5	66	2
3	M	No	Nobel	14	5	2.55	130	1
4	M	No	Nobel	3	1	3.1	78	1
5	F	No	Nobel	3	3	2.7	40	1
6	F	No	Nobel	5	4	3.2	80	2
7	F	No	Olympic	10	10	2.77	94	1
8	M	No	Olympic	13	8	3.3	77	1
9	F	No	Nobel	3	6	2.8	60	2
10	F	No	Nobel	12	1	3.7	94	8

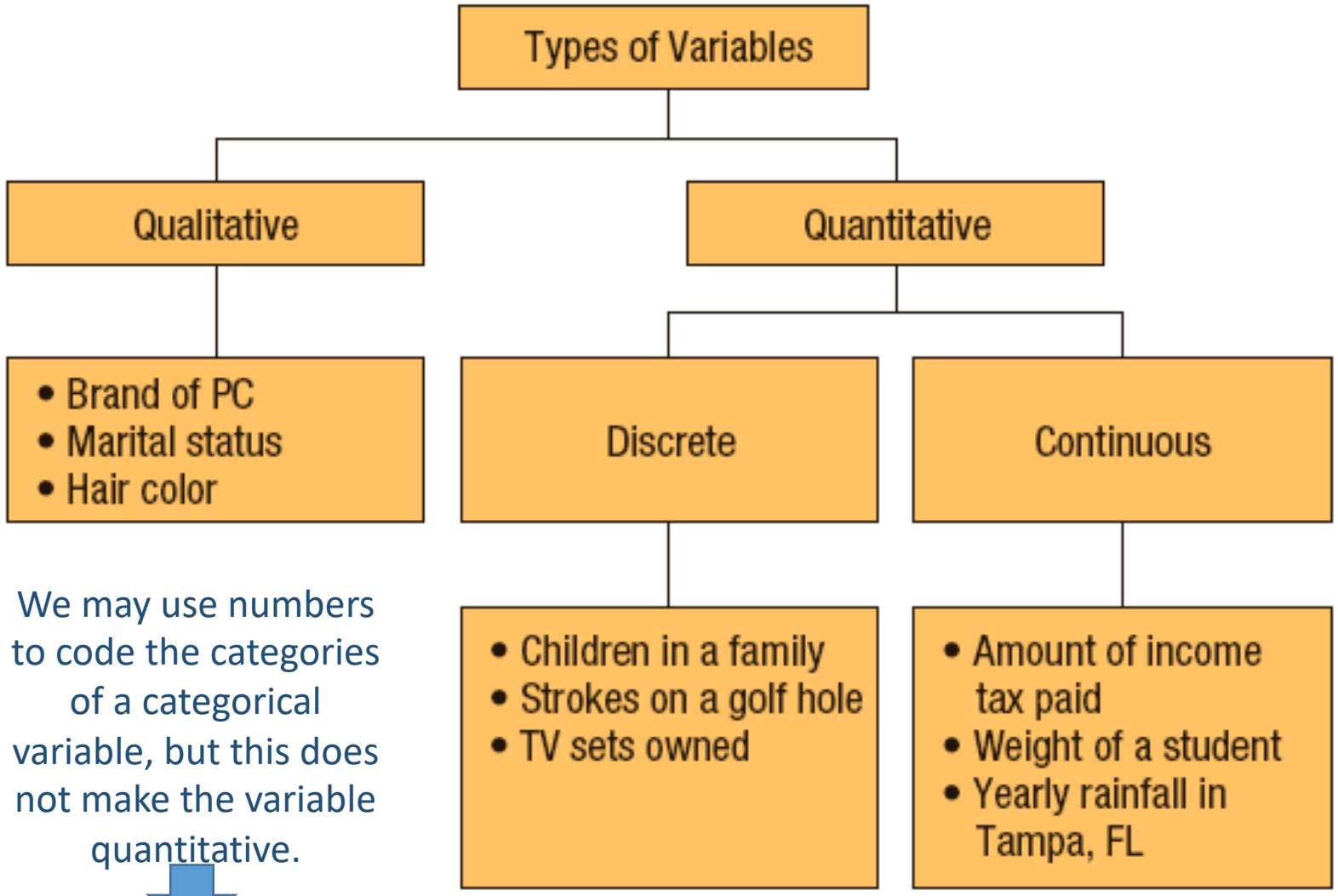
Example

- Student 1 is a male who does not smoke and who would prefer to win an Olympic gold medal over an Academy Award or a Nobel Prize.
- He says that he exercises 10 hours a week, watches television one hour a week, and that his grade point average is 3.13.
- His pulse rate was 54 beats per minute at the time of the survey, and he is the fourth oldest child in his family.

Exercise

- Student 2 is a Female who smoke and who would prefer to win Academy Award over Olympic gold medal or Nobel Prize.
- She says that She exercises 4 hours a week, watches television 7 hour a week, and that grade point average is 2.5.
- Her pulse rate was 66 beats per minute at the time of the survey, and is the 2nd child in her family.

Variables – Categorical and Quantitative Variables



- A categorical /qualitative variable divides the cases into ***groups***, placing each case into exactly one of two or more categories.
- A quantitative /numerical variable measures or records a numerical quantity for each case. Numerical operations like adding and averaging make sense for quantitative variables.

We may use numbers to code the categories of a categorical variable, but this does not make the variable quantitative.

For example, “gender" is categorical even if we choose to record the results as 1 for male and 2 for female.

Categorical and Quantitative Variables

ID	Gender	Smoke	Award	Exercise	TV	GPA	Pulse	Birth

- Note that the ID column is neither a quantitative nor a categorical variable. A dataset often has a column with names or ID numbers that are for reference only.
- Gender is categorical since it classifies students into the two categories of male and female.
- Smoke is categorical since it classifies students as smokers or nonsmokers.
- Award is categorical since students are classified depending on which award is preferred.
- Exercise, TV, GPA, and Pulse are all quantitative since each measures a numerical characteristic of each student. It makes sense to compute an average for each variable, such as an average number of hours of exercise a week.

Using Data to Answer a Question – single variable

In the student survey data, we might be interested in the following questions, each about a single variable:

- What percentage of students smoke?
- What is the average number of hours a week spent exercising?
- Are there students with unusually high or low pulse rates?
- Which award is the most desired?
- How does the average GPA of students in the survey compare to the average GPA of all students at this university?

Using Data to Answer a Question – between variables

- In the student survey data, we might ask the following questions about relationships between variables:
  - Who smokes more, males or females?
  - Do students who exercise more tend to prefer an Olympic gold medal? Do students who watch lots of television tend to prefer an Academy Award?
  - Do males or females watch more television?
  - Do students who exercise more tend to have lower pulse rates?
  - Do first-borns generally have higher grade point averages?
- These examples show that relationships might be between two categorical variables, two quantitative variables, or a quantitative and a categorical variable.

Example 1 -- Is there a “Sprinting Gene"?

- A gene called **ACTN3** encodes a protein which functions in fast-twitch muscles. Some people have a variant of this gene that cannot yield this protein.
- To address the question of whether this gene is associated with sprinting ability, geneticists tested people from three different groups: world-class sprinters, world-class marathon runners, and a control group of non-athletes.
- In the samples tested, 6% of the sprinters had the gene variant, compared with 18% of the non-athletes and 24% of the marathon runners. This study suggests that sprinters are less likely than non-sprinters to have the gene variant.

Example 1 -- Is there a “Sprinting Gene"?

- The cases are the people included in the study.
- One variable is whether the individual has the gene variant or not. Since we record simply “yes" or “no", this is a categorical variable.
- The second variable keeps track of the group to which the individual belongs. This is also a categorical variable, with three possible categories (sprinter, marathon runner, or non-athlete).
- We are interested in the relationship between these two categorical variables.
- The table of data must record answers for each of these variables and may or may not have an identifier column.

Name	Gene Variant	Group
Allan	Yes	Marathon runner
Beth	No	Sprinter
...	...	...

Example 2. Try to answer the questions

(a) What are the cases?

(b) What is the variable and is it quantitative or categorical?

- People in a city are asked if they support a new recycling law.
  - (a) The people who are asked
  - (b) Support the law or not; Categorical
- Collect data from a sample of teenagers with a question that asks “Do you eat at least five servings a day of fruits and vegetables?"
  - (a) The teenagers in the sample
  - (b) At least five servings or not; Categorical
- Estimate the bending strength of beams by bending 10 beams until they break and recording the force at which the beams broke
  - (a) The 10 beams
  - (b) Force at which each beam broke; Quantitative

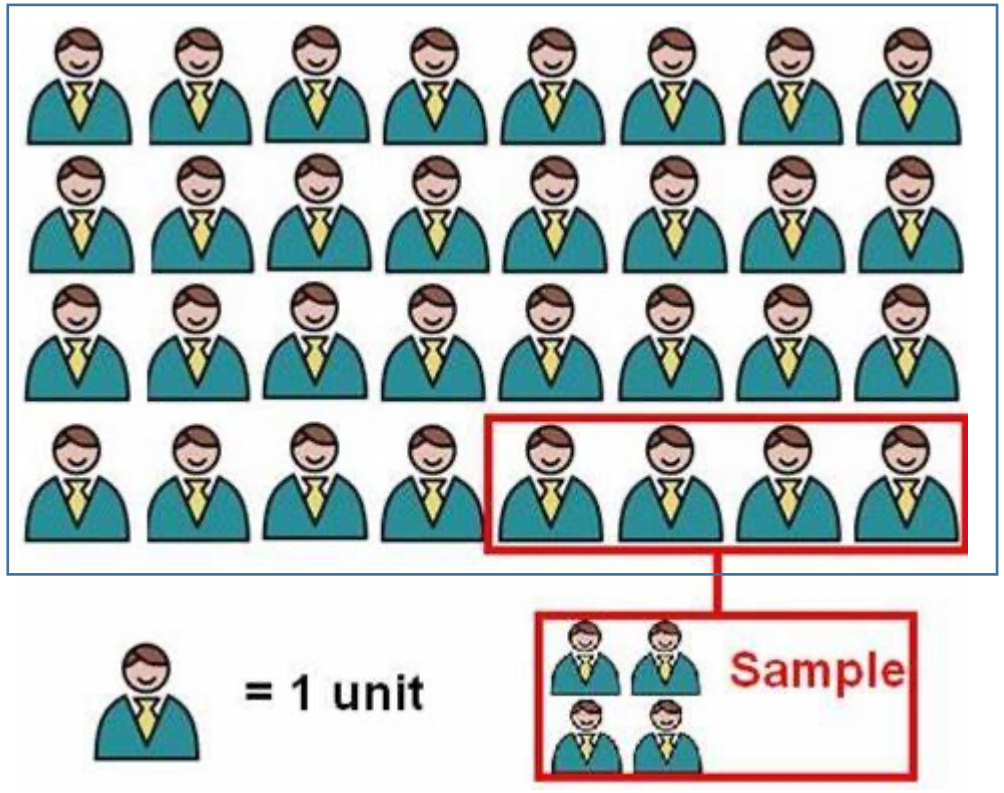
Outline

- ❑ Introduction to data
  - ❑ Applications
  - ❑ Concepts
- ❑ Collecting Data: Sampling
- ❑ Descriptive Statistics

Collecting data from population

- The US Census is conducted every 10 years and attempts to gather data about all people living in the US.
  - For example, the census shows that, for people living in the US who are at least 25 years old, 84.6% have at least a high school degree and 27.5% have at least a college bachelor's degree.
- The cases in the census dataset are all residents of the US, and there are many variables measured on these cases.
- The US census attempts to gather information from an entire population.

Collecting data from samples



- Usually, it is not feasible to gather data for an entire population (expensive and time consuming).
- In most circumstances, we can only work with a sample from what might be a very large population.

A population includes all individuals or objects of interest.

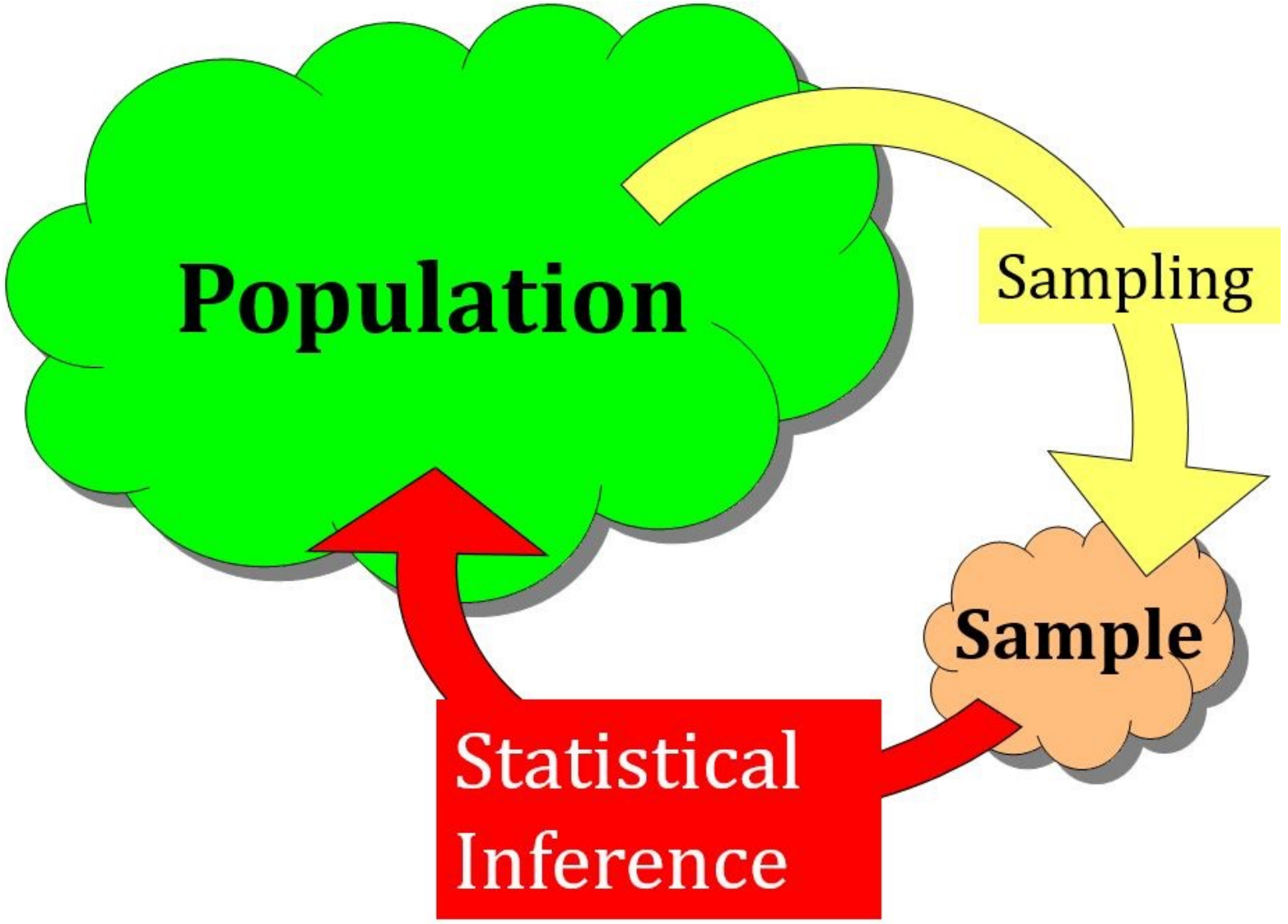
A Sample is all the cases that we have collected data on (a part of the population)

Example

- Aim: estimate what percent of people in the US wash their hands after using a public restroom
- Methods: researchers pretended to comb their hair while observing 6,000 people in public restrooms throughout the United States
- Results
  - 85% of the people who were observed washed their hands after going to the bathroom.
  - The sample is the 6,000 people who were observed.
  - A reasonable population to generalize to would be all people in the US who use public restrooms.

Statistical inference

- Statistical inference is the process of using data from a sample to gain information about the population.
  - Selecting a sample from the population is critical because the process used to collect the sample determines whether valid inference is even possible.

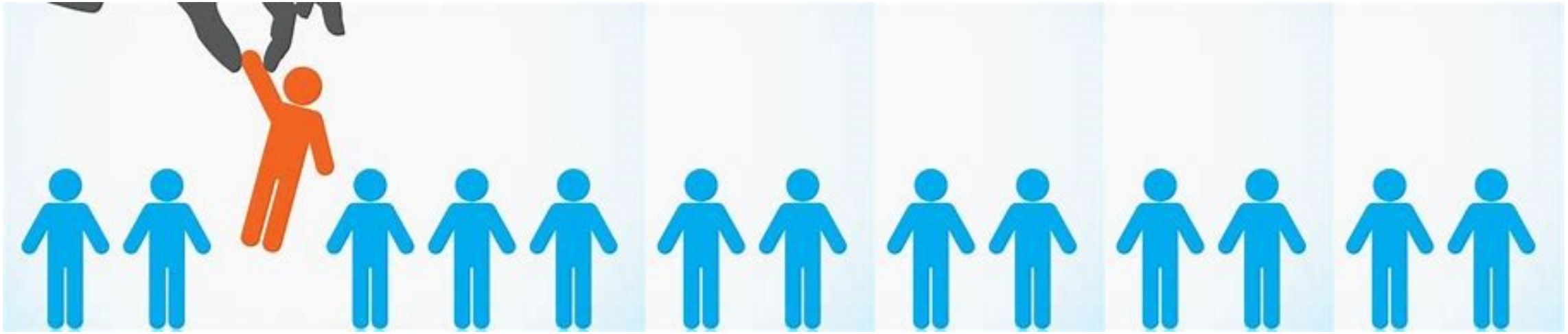


Example: Dewey defeats Truman?

- The paper was published before the conclusion of the 1948 presidential election of US, and was based on the results of a large telephone poll
- However, Harry S. Truman won the election
- What went wrong?

Example: Dewey defeats Truman?

- Sample: all the people who participated in the telephone poll
- Population: all voting Americans
- The pollsters wanted to estimate the percentage of all voting Americans who would vote for each candidate.
- One reason the telephone poll may have yielded inaccurate results is that people with telephones in 1948 were not representative of all American voters.
  - People with telephones tended to be wealthier and prefer Dewey while people without phones tended to prefer Truman.



Sampling bias

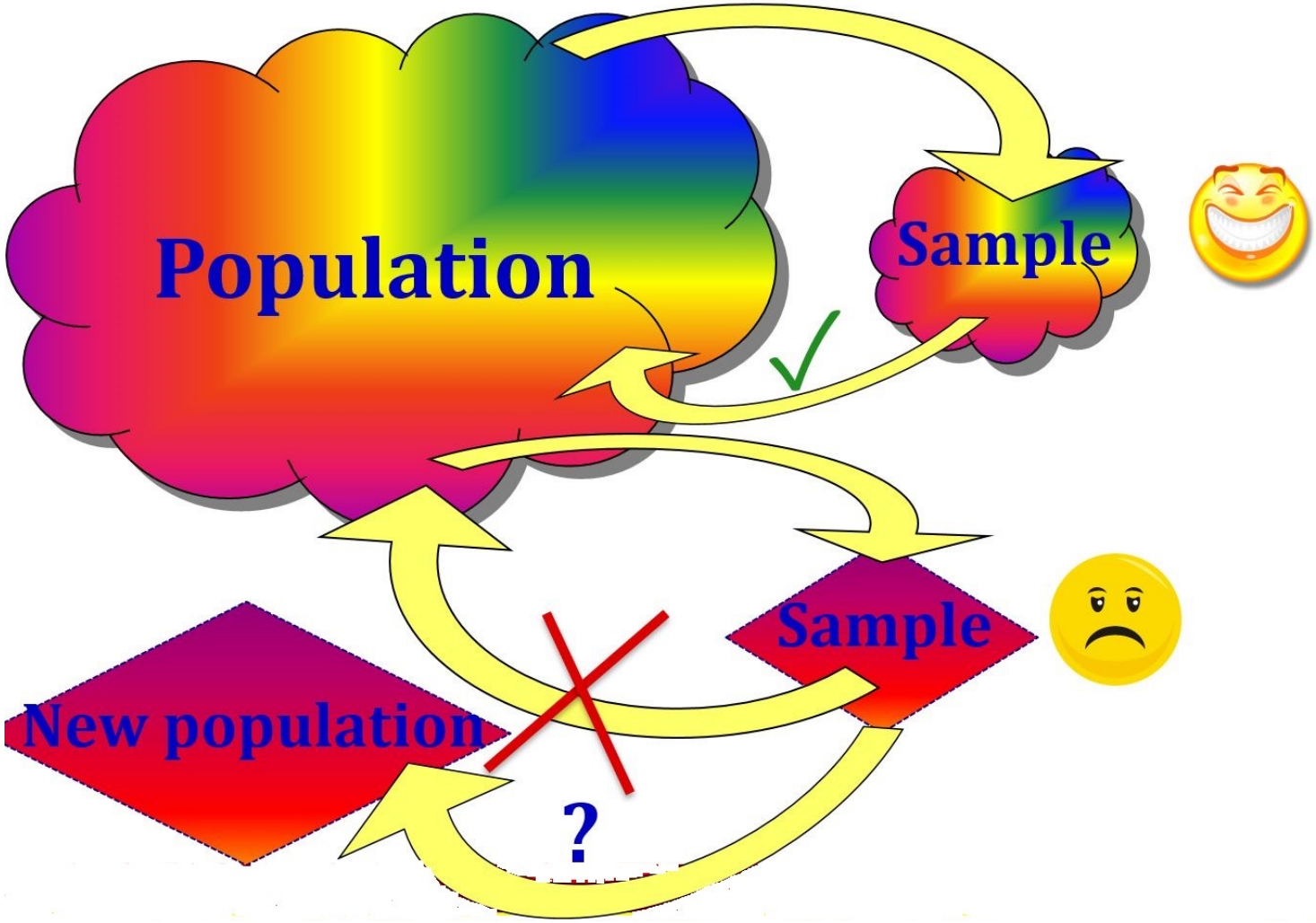
- Sampling bias occurs when the method of selecting a sample causes the sample to differ from the population in some relevant way.
- If sampling bias exists, we cannot trust generalizations from the sample to the population

Example:If you ask people in a survey about how often they shower, or how often they recycle, your data is going to be biased by the fact that nobody wants to admit to doing something that is considered socially undesirable.

Adding in a sampling bias into your data collection is an important tool if you want to lie, cheat, manipulate, or mislead with your study results!

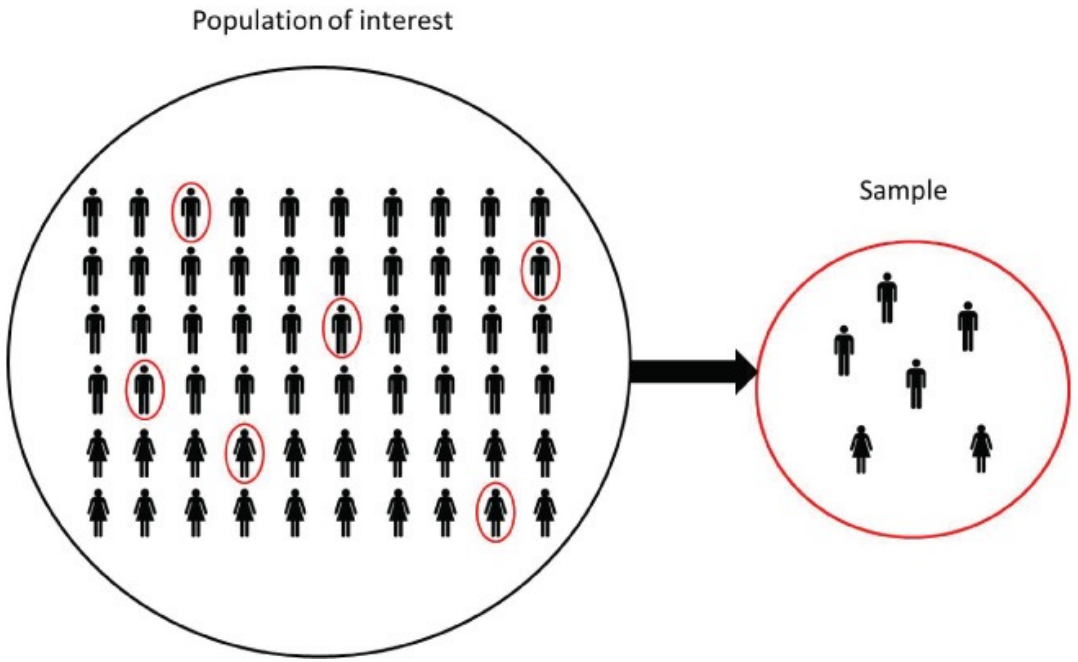
Sampling

Inference can only extend to a population that “looks like” the sample.



Random Sampling

- How do we get a sample that looks like the population?
- The key is random sampling!!!
- A random sample will resemble the population!
- Random sampling avoids sampling bias!



Example: Sampling the soup

- Think of tasting a bowl of soup...
  - Population = entire bowl of soup
  - Sample = whatever is in your tasting bites
- If you take bites non-randomly from the soup (if you stab with a fork, or prefer noodles to vegetables), you may not get a very accurate representation of the soup
- If you take bites at random, only a few bites can give you a very good idea for the overall taste of the soup

Simple Random Sample

- In a simple random sample, each unit of the population has the same chance of being selected, regardless of the other units chosen for the sample
- Like drawing names out of a hat

Example: Election Polling

- Right before the 2008 presidential election of US, the Gallup Poll randomly sampled and collected data on 2,847 Americans. Of those sampled, 52% supported Barack Obama and 42% supported John McCain.
- Can we generalize these results to the entire population of 129 million voters in order to estimate the popular vote in the election?
- Yes! In the actual election, 53% voted for Obama and 46% voted for McCain.
- Amazingly, if a simple random sample is selected, even a small sample can yield valid inferences for all 129 million American voters!

Outline

- ❑ Introduction to data
  - ❑ Applications
  - ❑ Concepts
- ❑ Collecting Data: Sampling
- ❑ Descriptive Statistics

Descriptive Statistics

Which Group is Smarter?

Class A--IQs of 13 students: 102, 128, 131, 98, 140, 93, 110, 115, 109, 89, 106, 119, 97

Class B--IQs of 13 students: 127, 96, 131, 80, 120, 109, 162, 111, 109, 103, 87, 105, 93

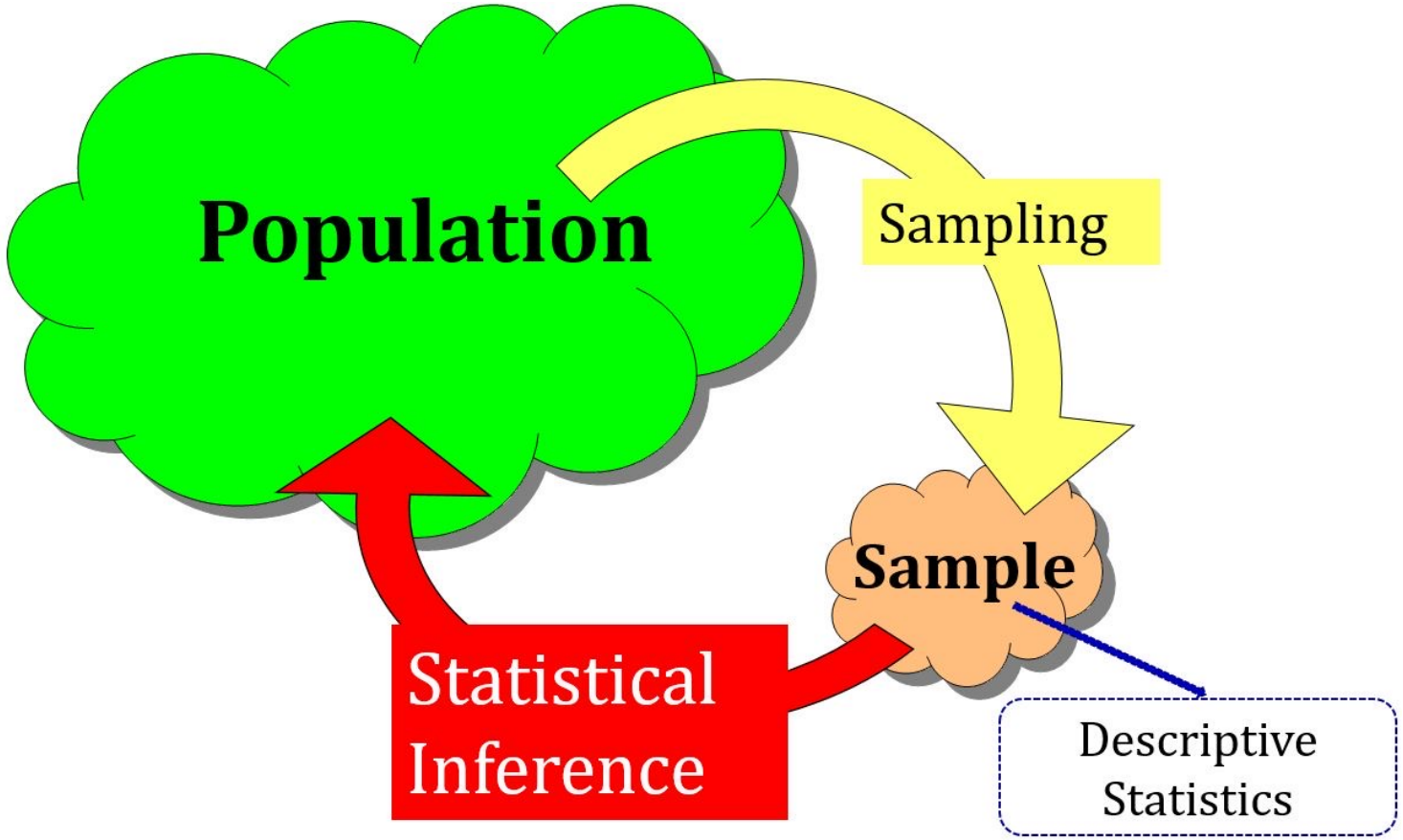
Each individual may be different. If you try to remember the value of each member, you will become overwhelmed and fail to have a overview of the group.

Class A--Average IQ=110.54 ⟷ They are roughly the same! ⟷ Class B--Average IQ=110.23

With a summary descriptive statistic, it is much easier to answer our question.

Descriptive statistics

- In order to make sense of data, we need ways to summarize and visualize it
- Summarizing and visualizing variables and relationships between two variables is often known as descriptive statistics (also known as exploratory data analysis)
- Type of summary statistics and visualization methods depend on the type of variable(s) being analyzed (categorical or quantitative)



Application1--Is there one true love for each person?

- A telephone survey asked 2,625 adults ages 18 and older

- "Some people say there is only one true love for each person. Do you agree or disagree?" The survey participants were selected randomly, by landlines and cell phones.

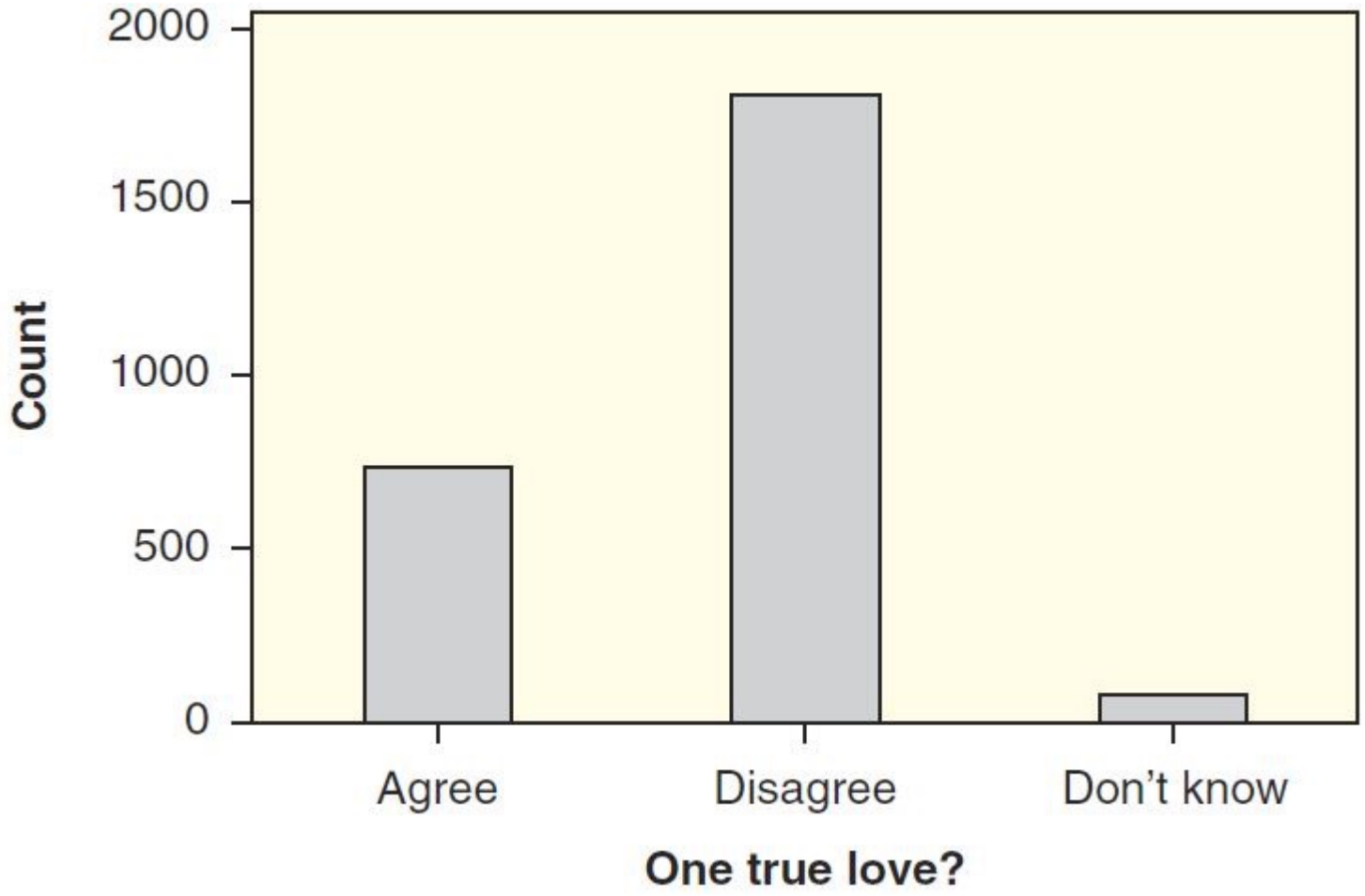
  - The sample is the 2,625 people who were surveyed.
  - The population is all US adults ages 18 or older who have a landline telephone or cell phone.
    - Since the sampling was random, there is no sampling bias, so the sample results can probably generalize to the population.
  - The cases are the adults who answered the survey questions.
  - The description indicates that there are two variables.
    - One variable is whether the responder agrees or disagrees with the statement that each person has only one true love.
    - A second variable is gender. Both variables are categorical.

Application1--Is there one true love for each person?

A frequency table shows the number of cases that fall in each category

Response	Frequency
Agree	735
Disagree	1812
Don't know	78
Total	2625

In a bar chart, the height of the bar is the number of cases falling in each category



Proportion

- The proportion in some category is found by

$$\text{Proportion in a category} = \frac{\text{Number in that category}}{\text{Total number}}$$

- We use different notation to indicate whether a quantity such as a proportion comes from a sample or an entire population
  - The proportion for a sample is denoted $\hat{p}$ and read "p-hat".
  - The proportion for a population is denoted p.
- Question: What proportion of the responders agree with the statement that we all have exactly one true love?

$$\hat{p} = \frac{\text{Number who agree}}{\text{Total number}} = \frac{735}{2625} = 0.28$$

Answer: The proportion who agree that there is one true love is 0.28 or 28%

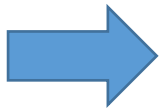
Response	Frequency
Agree	735
Disagree	1812
Don't know	78
Total	2625

Relative Frequency Table

- A relative frequency table shows the proportion of cases that fall in each category. All the numbers in a relative frequency table sum to 1.

Response	Frequency
Agree	735
Disagree	1812
Don't know	78
Total	2625

frequency table

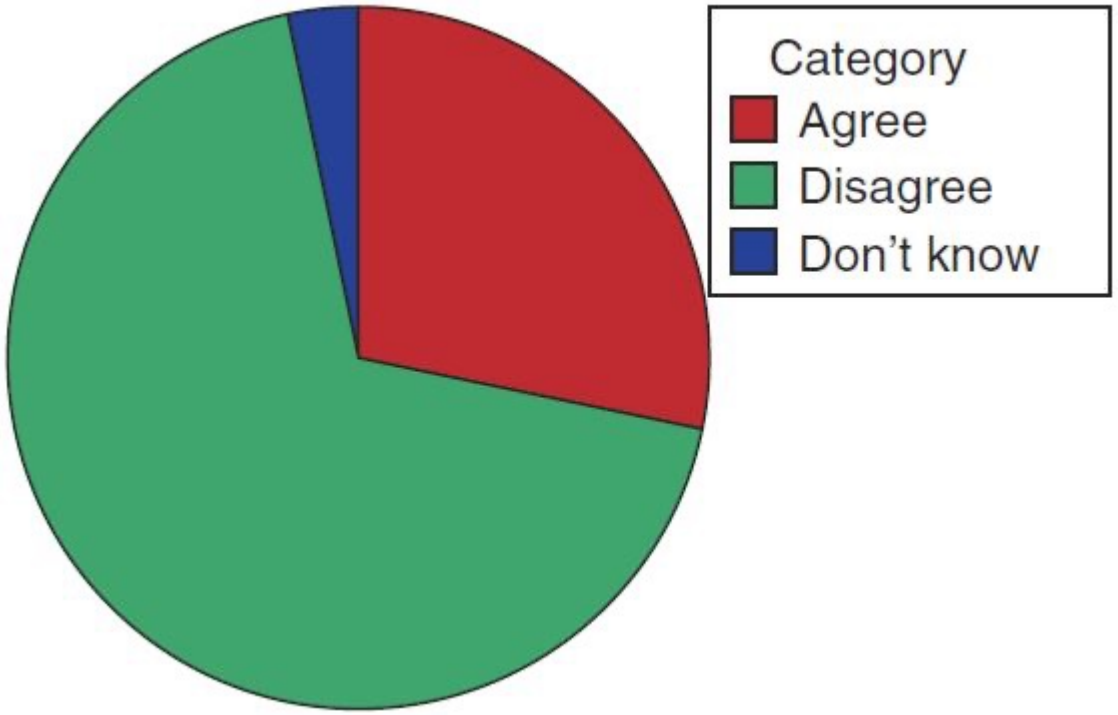


Response	Frequency
Agree	0.28
Disagree	0.69
Don't know	0.03
Total	1.00

relative frequency table

Pie Chart

- In a pie chart, the relative area of each slice of the pie corresponds to the proportion in each category



Application1--Is there one true love for each person?

- In addition to finding out the proportion who agree with the statement, they also wanted to find out if the proportion who agree is different between males and females.
- The question is asking about a relationship between two categorical variables.
- To investigate a possible relationship between two categorical variables we use a two-way table.

Two-Way Table

- The categories for one variable are listed down the side (rows) and the categories for the second variable are listed across the top (columns).
- Each cell of the table contains the count of the number of data cases that are in both the row and column categories.

Response	Frequency
Agree	735
Disagree	1812
Don't know	78
Total	2625

frequency table

Response	Male	Female	Total
Agree	372	363	735
Disagree	807	1005	1812
Don't know	34	44	78
Total	1213	1412	2625

two-way table

Are men or women more hopelessly romantic?

$$\text{Proportion of females} = \frac{\text{Number of females}}{\text{Total number}} = \frac{1412}{2625} = 0.54$$

$$\text{Proportion of females who agree} = \frac{\text{Number of females who agree}}{\text{Number of females}} = \frac{363}{1412} = 0.26$$

$$\text{Proportion who agree that are female} = \frac{\text{Number of females who agree}}{\text{Number of who agree}} = \frac{363}{735} = 0.49$$

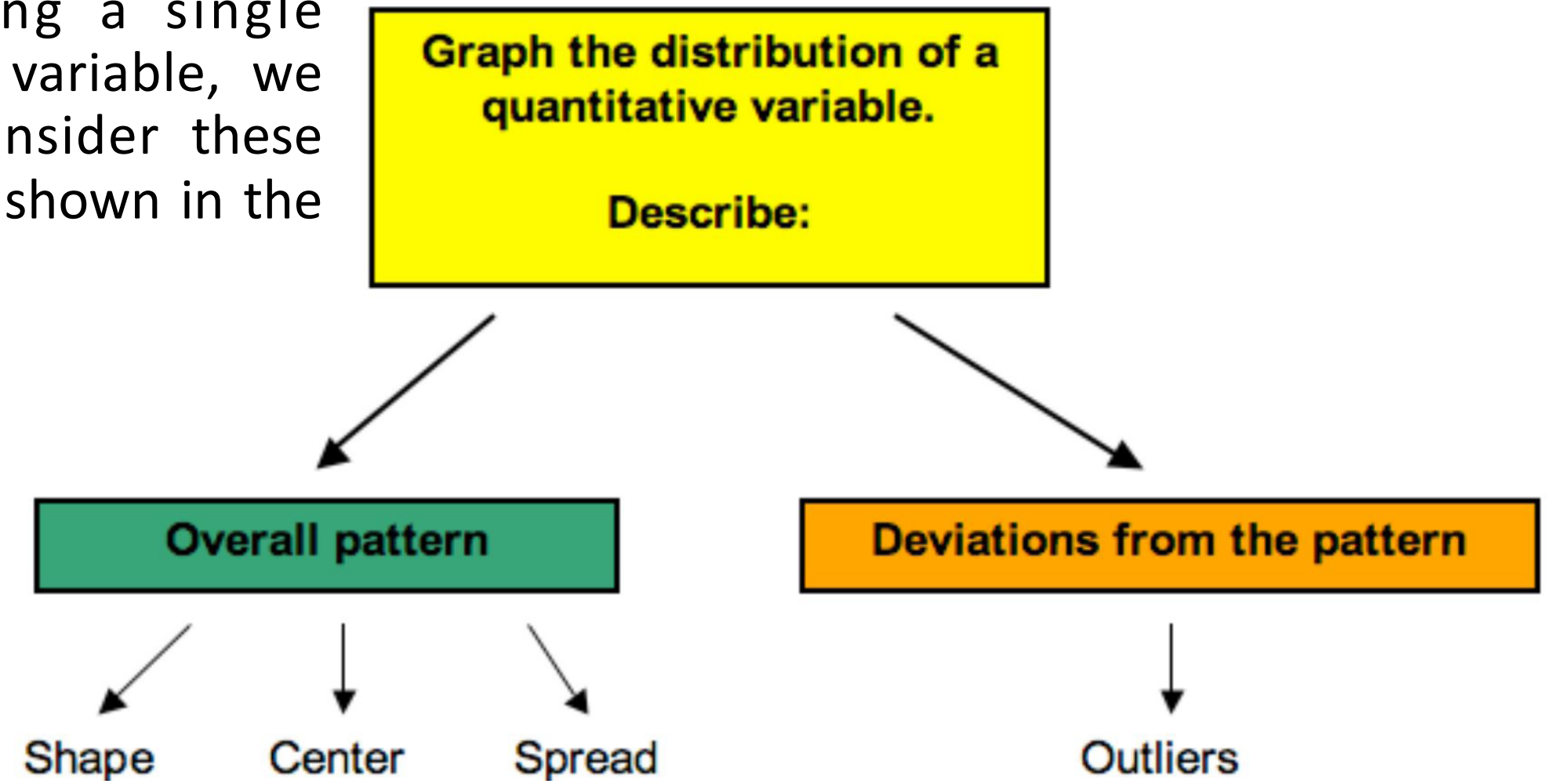
They are different!

$$\text{Proportion of males who agree} = \frac{\text{Number of males who agree}}{\text{Number of males}} = \frac{372}{1213} = 0.31$$

- 31% of the males in the survey agree that there is one true love for each person while only 26% of the females agree with that statement. In this sample, males are more likely than females to believe in one true love.

Distribution

- In describing a single quantitative variable, we generally consider these questions as shown in the right graph.



Example -- How long does an elephant live?

Some data on the typical lifespan for several mammals

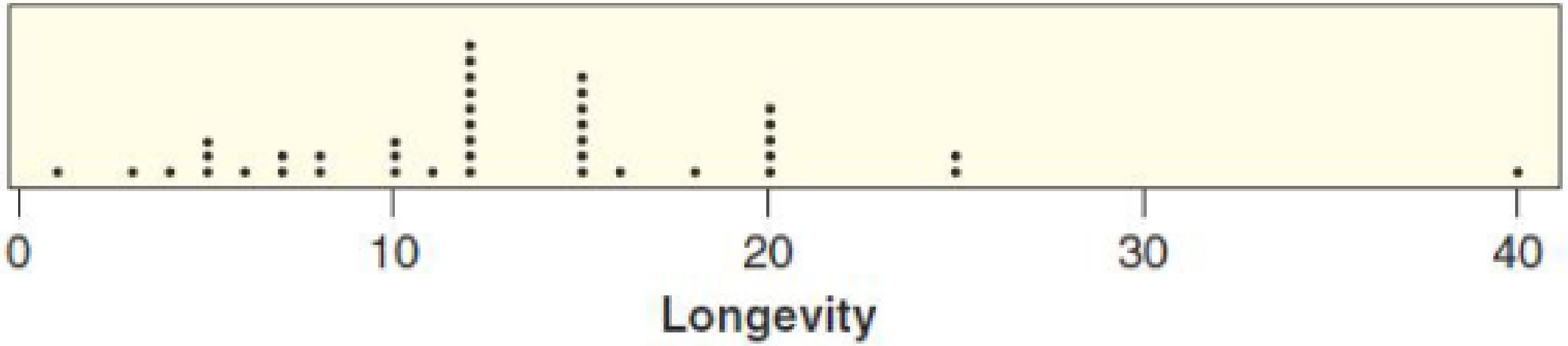
Species	Longevity	Species	Longevity	Species	Longevity
Baboon	20	Elephant	40	Mouse	3
Black bear	18	Elk	15	Opossum	1
Grizzly bear	25	Fox	7	Pig	10
Polar bear	20	Giraffe	10	Puma	12
Beaver	5	Goat	8	Rabbit	5
Buffalo	15	Gorilla	20	Rhinoceros	15
Camel	12	Guinea pig	4	Sea lion	12
Cat	12	Hippopotamus	25	Sheep	12
Chimpanzee	20	Horse	20	Squirrel	10
Chipmunk	6	Kangaroo	7	Tiger	16
Cow	15	Leopard	12	Wolf	5
Deer	8	Lion	15	Zebra	15
Dog	12	Monkey	15		
Donkey	12	Moose	12		

A graph is useful to help us visualize the shape of a distribution.

How to describe these data in a graph?

Graph type 1--Dotplot

- A common way to visualize the shape of a moderately sized dataset is a dotplot.



- All but one typically live between 1 and 25 years, while the elephant's lifespan of 40 years is much higher than the rest.
- The value of 40 years appears to be an outlier for longevity in this group of mammals.

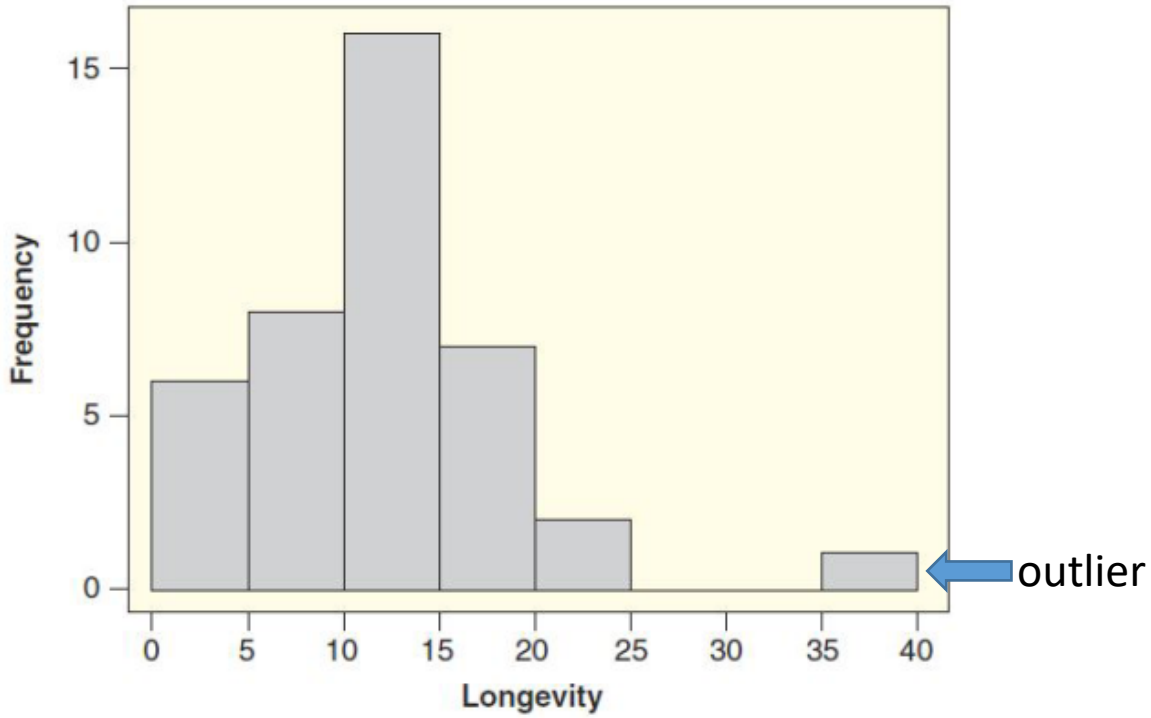
Graph type 2--histogram

- Step 1: group the longevity data into five-year intervals (called bins), obtain the frequency table
- Step 2: plot histogram. The height of the each bar corresponds to the number of cases within that range of the variable

Longevity (years)	Frequency Count
1–5	6
6–10	8
11–15	16
16–20	7
21–25	2
26–30	0
31–35	0
36–40	1
Total	40

outlier

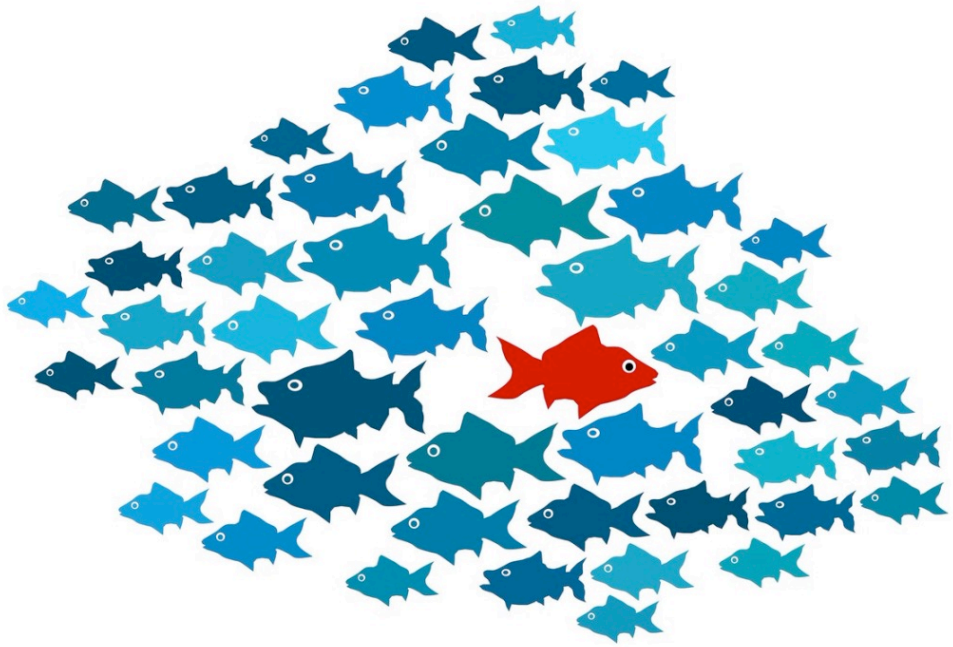
Frequency table



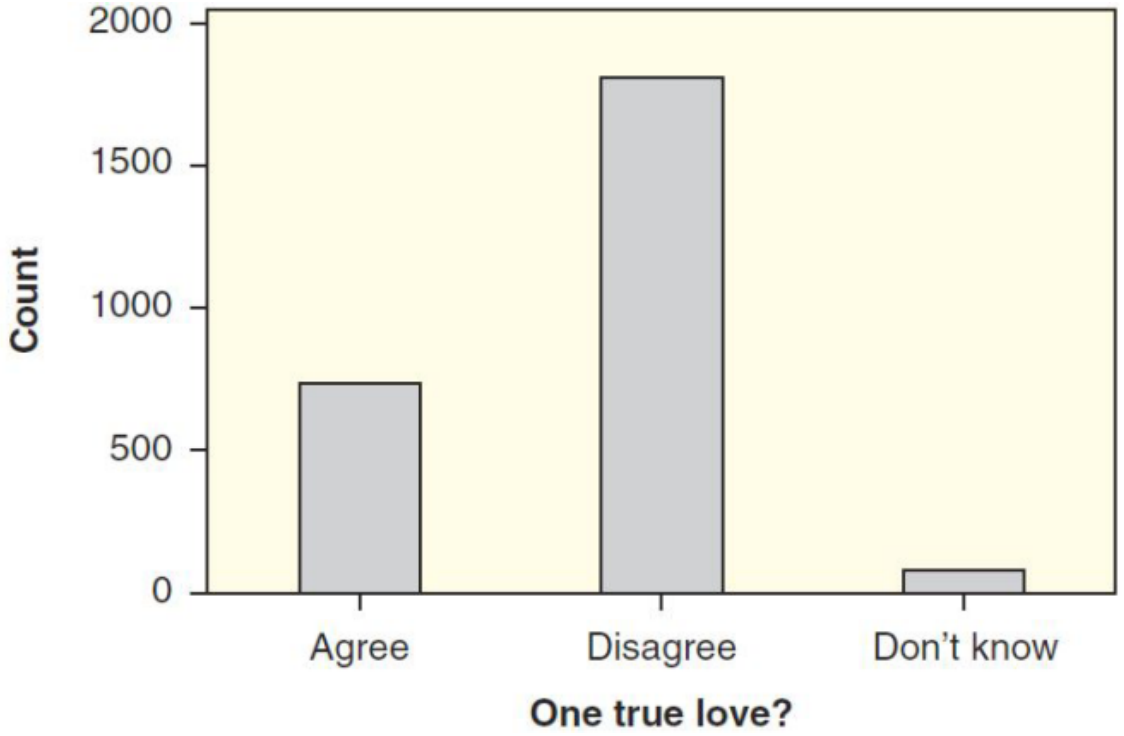
Histogram

Outlier

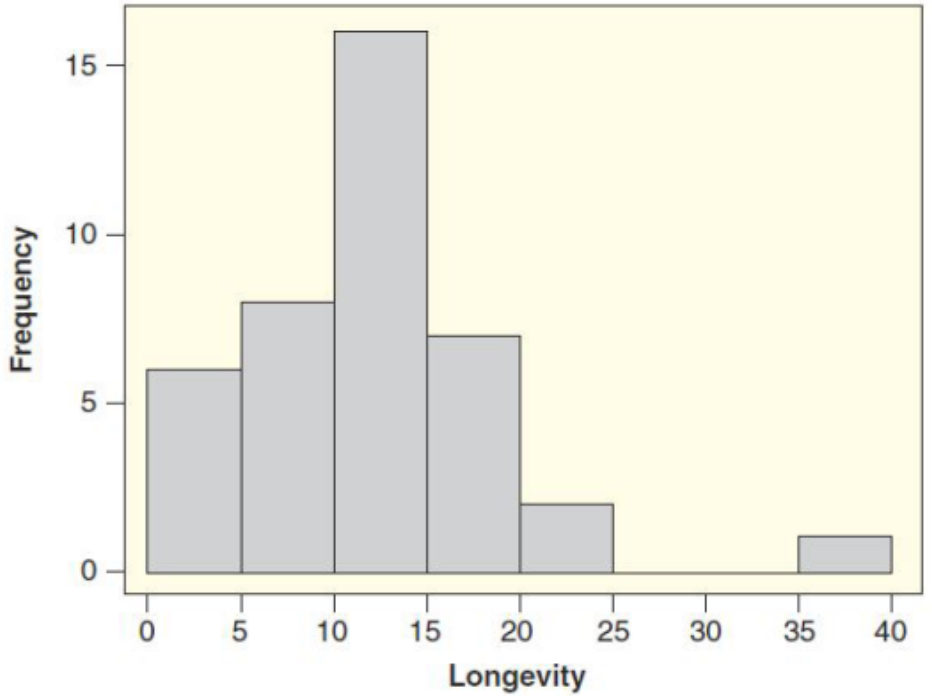
- An outlier is an observed value that is notably distinct from the other values in a dataset. Usually, an outlier is much larger or much smaller than the rest of the data values.



Review: Bar Chart vs. Histogram



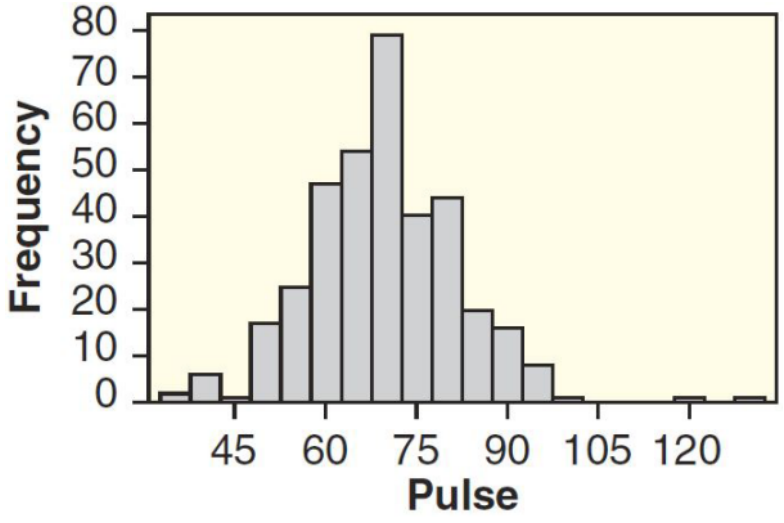
- A bar chart is for categorical data
- The number of bars equals the number of categories, and the number in each category is fixed
- The x-axis has no numerical scale



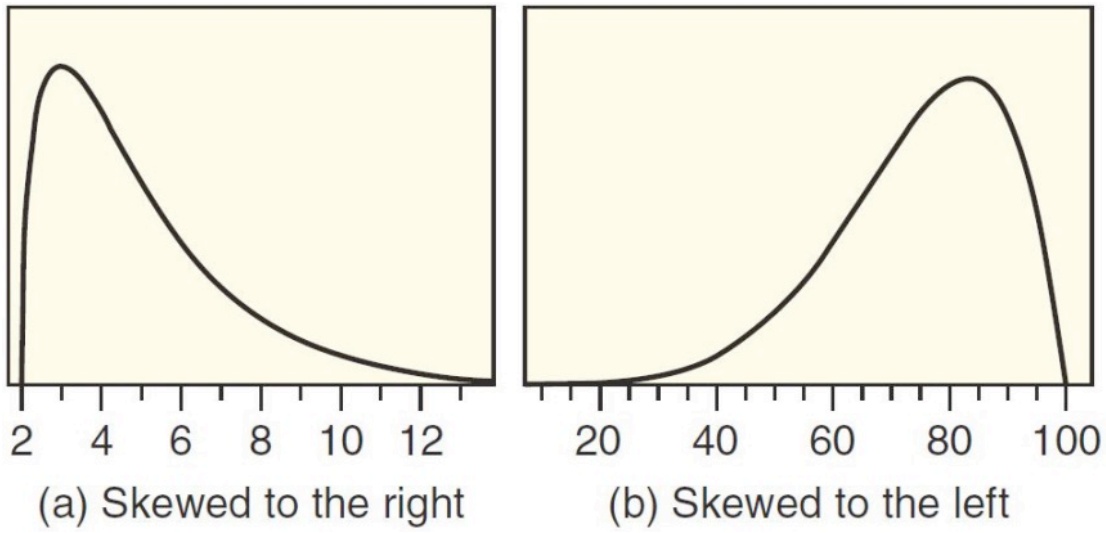
- A histogram is for quantitative data
- The number of bars in a histogram is up to you, the appearance can differ with different number of bars
- The x-axis includes numerical values

Distribution--Shape

- A distribution is considered symmetric if we can fold the plot over a vertical center line and the two sides match closely.



Other than the outliers, this histogram is quite symmetric.

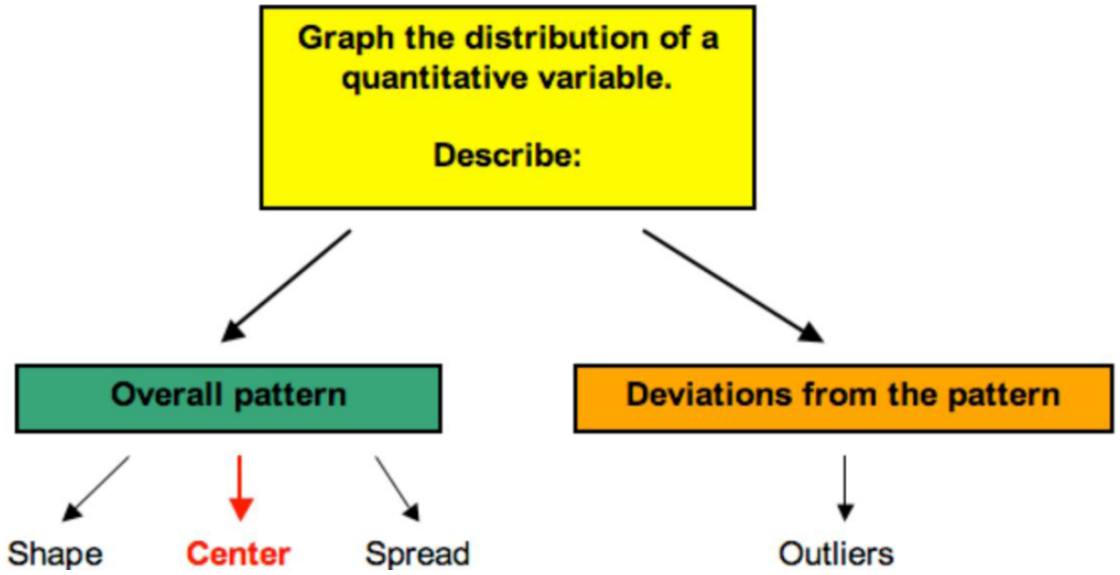


(a) Skewed to the right

(b) Skewed to the left

Distribution--center

- We can also summarize important features of a distribution numerically.
- Two summary statistics that describe the center or location of a distribution for a single quantitative variable are the **mean** and the **median**.



Example -- How long does an elephant live?

Some data on the typical lifespan for several mammals

Species	Longevity	Species	Longevity	Species	Longevity
Baboon	20	Elephant	40	Mouse	3
Black bear	18	Elk	15	Opossum	1
Grizzly bear	25	Fox	7	Pig	10
Polar bear	20	Giraffe	10	Puma	12
Beaver	5	Goat	8	Rabbit	5
Buffalo	15	Gorilla	20	Rhinoceros	15
Camel	12	Guinea pig	4	Sea lion	12
Cat	12	Hippopotamus	25	Sheep	12
Chimpanzee	20	Horse	20	Squirrel	10
Chipmunk	6	Kangaroo	7	Tiger	16
Cow	15	Leopard	12	Wolf	5
Deer	8	Lion	15	Zebra	15
Dog	12	Monkey	15		
Donkey	12	Moose	12		

A statistical measure is useful to help us summarize the center of a distribution.

- The sample size, the number of cases in the sample, is denoted by n
- We often let x or y stand for any variable, and x_1, x_2, ..., x_n represent the n values of the variable x
- $x_1 = 20$, $x_2 = 18$, $x_3 = 25$, ...

Mean

- The mean for a single quantitative variable is the numerical average of the data values:

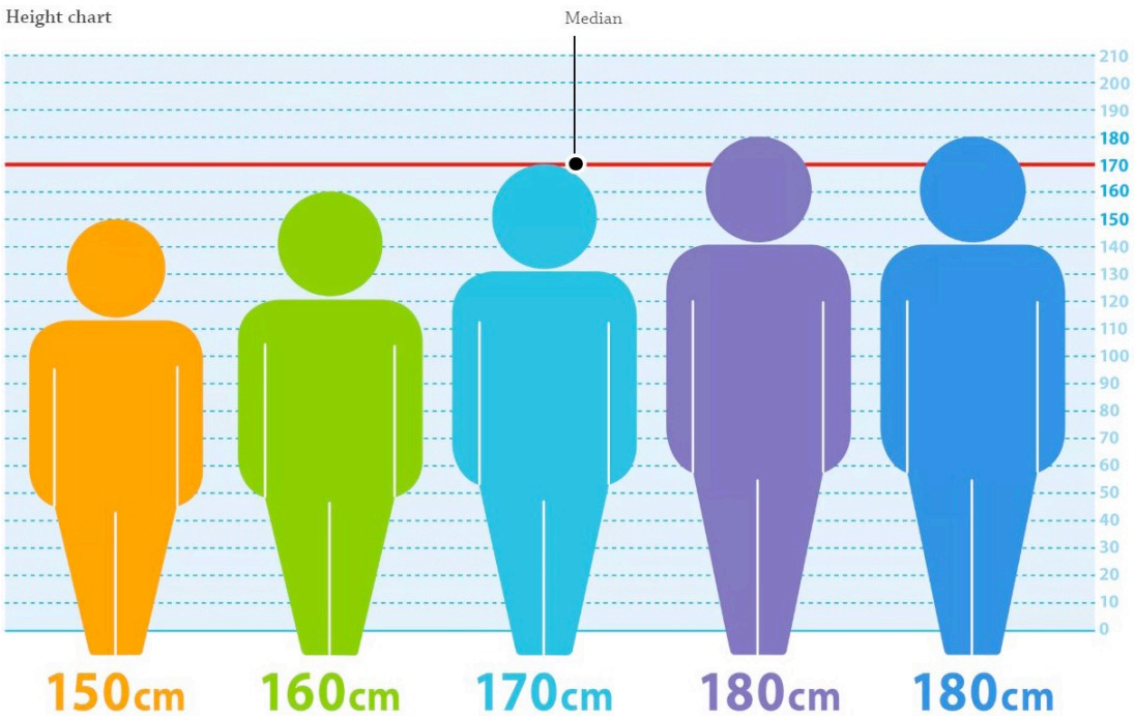
$$\text{Mean} = \frac{\text{Sum of all data values}}{\text{Number of data values}} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\Sigma x}{n}$$

- The Greek letter Σ (read “sigma”) is used to add all the values of x.
- The longevity data yield a sum of Σx = 526 years and thus the mean longevity for this sample of 40 mammals is 526/40 = 13.15 years.

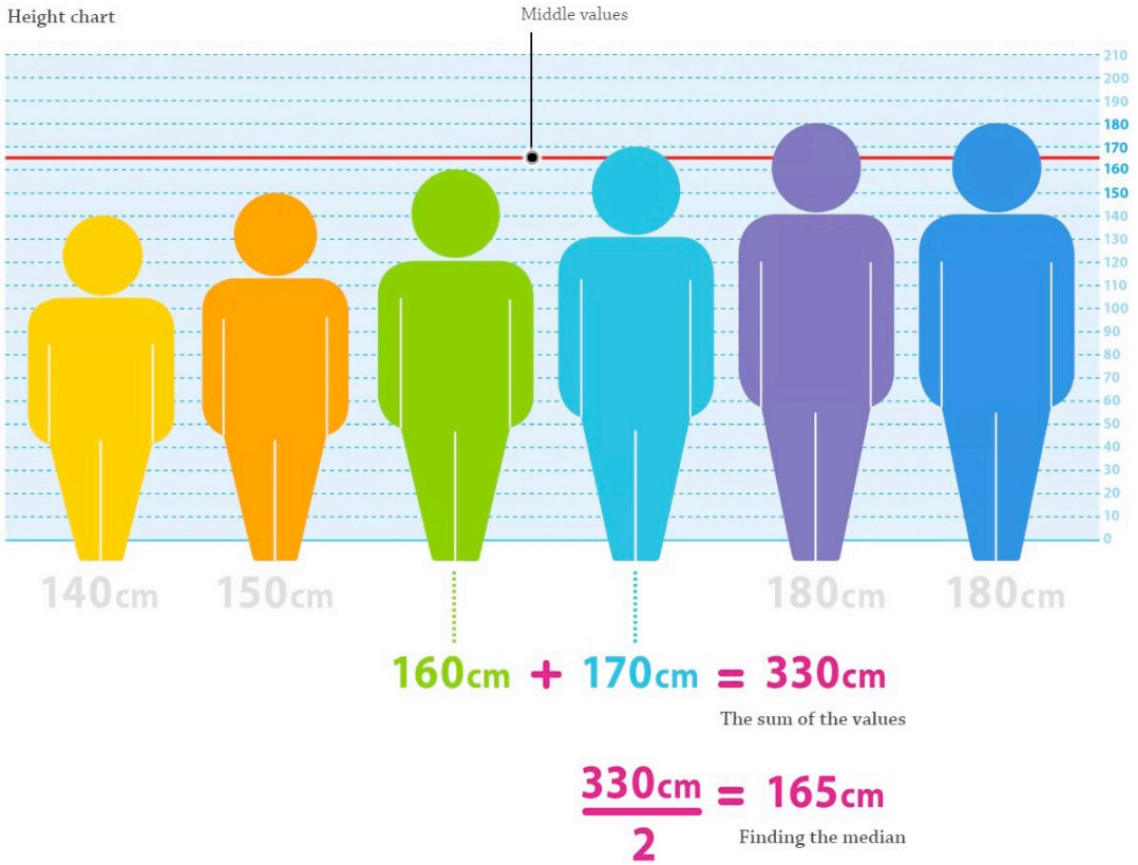
Mean--notation and example

- The mean of a sample is denoted $\bar{x}$ and read "x-bar"
  - For a random sample of 50 seniors from a high school, the average SAT score was 582 on the Math portion of the test. The mean of 582 represents the mean of a sample, so we use the notation $\bar{x}$ for the mean, and we have $\bar{x}$ = 582.
- The mean of a population is denoted μ, which is the Greek letter "mu"
  - Nearly 1.6 million students in the class of 2010 took the SAT, and the average score overall on the Math portion was 516. The mean of 516 represents the mean for everyone who took the exam in the class of 2010, so we use the notation μ for the population mean, and we have μ = 516.

Median: splits the data in half



The median, denoted m, is the middle value when the data are ordered.



If there are an even number of values, the median is the average of the two middle values.

Median -- example

Find the median and the mean for the heart rates of 20-year-old patients and 55-year-old patients.

20-year-old patients:	108,	68,	80,	83,	72			
55-year-old patients:	86,	86,	92,	100,	112,	116,	136,	140

- For the 20-year-old patients, we put the heart rates in order from smallest to largest:

68, 72, 80, 83, 108

- The middle number is the third largest value, so the median hart rate m = 80.
- These values are from a sample, so we use the

$$\text{Mean} = \bar{x} = \frac{108 + 68 + 80 + 83 + 72}{5} = \frac{411}{5} = 82.2$$

Median -- example

- For the 55-year-old patients, there are an even number (n = 8) of values in this dataset, there is no single number in the middle.
  - The median is the average of the two middle numbers:

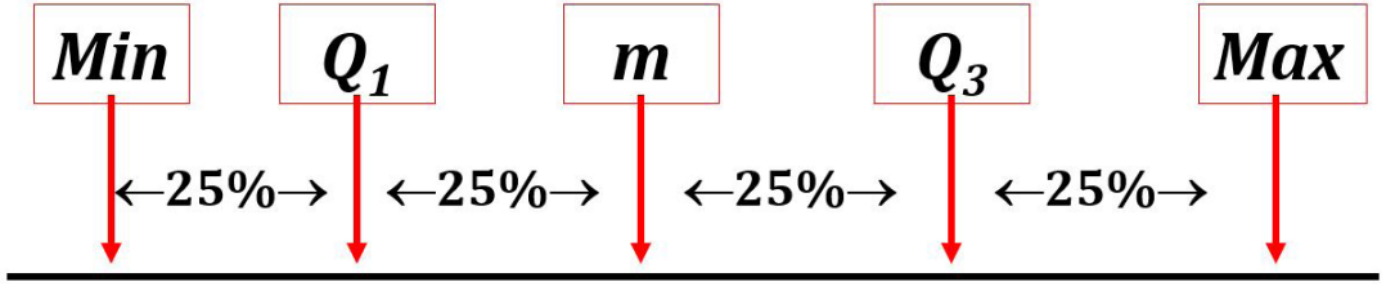
$$m = \frac{100 + 112}{2} = 106$$

  - The mean of the heart rates for the 55-year-old patients is

$$\bar{x} = \frac{868}{8} = 108.5$$

Other measures of location and spread

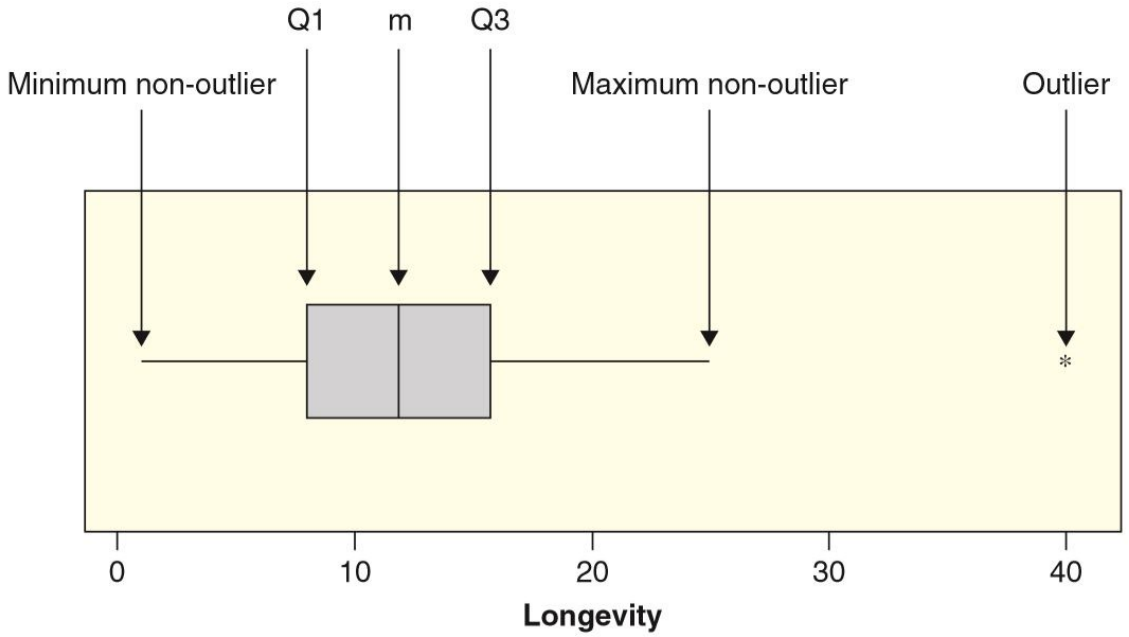
- Maximum = the largest value
- Minimum = the smallest value
- Mode: The value that has the highest frequency.
- Quartiles:
  - Q_1 = median of the values below m
  - Q_3 = median of the values above m
- Five Number Summary



- Range = Max - Min; Is the range resistant to outliers? No!
- Interquartile Range (IQR) = Q_3-Q_1; Is the IQR resistant to outliers? Yes!

Example--Boxplot of longevity of mammals

- The five number summary for mammal longevities is (1,8,12,16,40). We have $Q_1 = 8$ and $Q_3 = 16$ so the interquartile range is IQR =16-8= 8
- We compute Q1-1.5(IQR) = 8-1.5(8) = 8-12 =-4 and $Q_3 + 1.5(IQR) = 16 + 1.5(8) = 16 + 12 = 28$



- The box itself spans the range from the **25th percentile** to the **75th percentile** and the whiskers go out to the most extreme data point that does not exceed a certain bound (1.5 times the interquartile range by default).
- Outliers, observations that are "suspiciously" distant from the rest of the data, are plotted as a dot.

We adopt Method2 in this course

Method 2

1. Use the median to divide the ordered data set into two-halves.
    - If there are an odd number of data points in the original ordered data set, **include** the median (the central value in the ordered list) in both halves.
    - If there are an even number of data points in the original ordered data set, split this data set exactly in half.
2. The lower quartile value is the median of the lower half of the data. The upper quartile value is the median of the upper half of the data.

The values found by this method are also known as "Tukey's hinges";[4] see also midhinge.

https://en.m.wikipedia.org/wiki/Quartile

Example for odd number of data point

- Ordered Data Set: 6, 7, 15, 36, 39, 40, 41, 42, 43, 47, 49

Q1=25.5, m=40, Q3=42.5

Example for even number of data point

- Ordered Data Set: 7, 15, 36, 39, 40, 41

7, 15, 36

Median=15

36,39

Median=(36+39)/2=37.5

39, 40, 41

Median=40

Q1=15, m=37.5, Q3=40

Exercise

DataSet1	1 3 5 6 9 11 12 13 19 21 22 32
DataSet2	6 9 11 12 13 19 21 22 32

Even number

DataSet1 1 3 5 6 9 11 12 13 19 21 22 32

1 3 5 6 9 11 12 13 19 21 22 32

Q1=(5+6)/2=5.5
M=(11+12)/2=11.5
Q3=(19+21)/2=20

Odd number

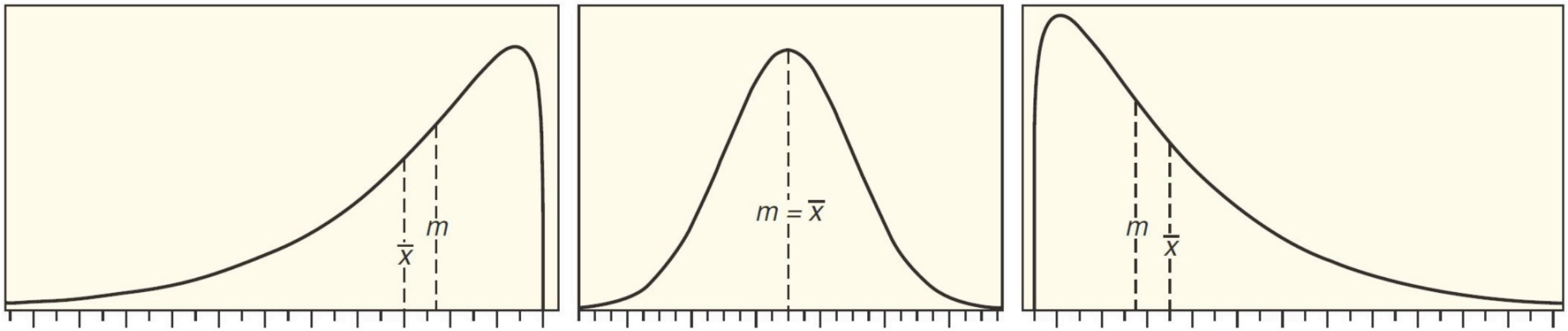
DataSet2 6 9 11 12 13 19 21 22 32

6 9 11 12 13 13 19 21 22 32

Q1=11
M=13
Q3=21

Mean and median for different shaped distributions

- For symmetric distributions, the mean and the median will be the same
- For skewed distributions, the mean will be more pulled towards the direction of skewness



(a) Mean < Median (b) Mean = Median (c) Mean > Median

Resistance

A statistic is resistant if it is relatively unaffected by extreme values. The median is resistant, while the mean is not.

20-year-old patients:	108,	68,	80,	83,	72

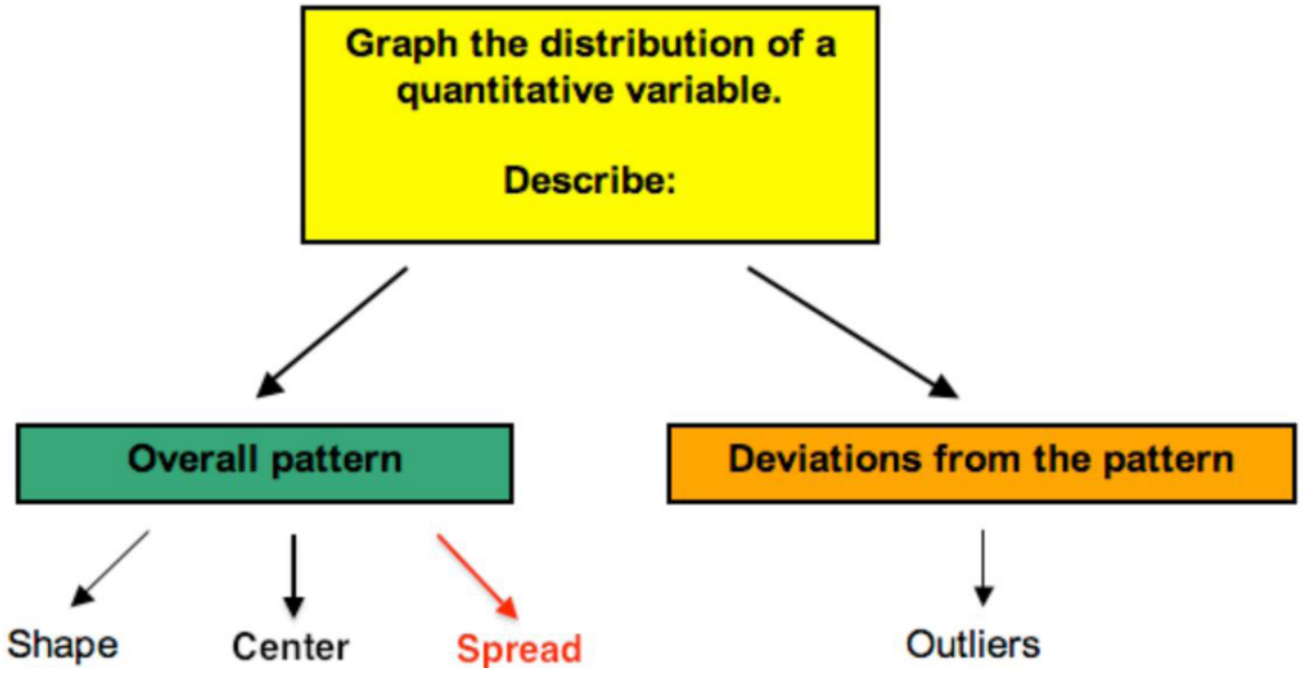
- The mean and the median heart rate for n = 5 patients in their twenties are given by $\bar{x}$ = 82.2 and m = 80.
- Suppose that the patient with a heart rate of 108 instead had an extremely high heart rate of 200.
- The median doesn't change at all, since 80 is still the middle value. The mean increases to $\bar{x}$ = 100.6.
- The extreme value of 200 has a large effect on the mean but little effect on the median.

Which measure should we use? Mean or median?

- Wealth per adult
  - Mean wealth per adult in Hong Kong: 1.445 million HKD
  - Median wealth per adult in Hong Kong: 268,000 HKD
- Why are them so different?
  - Median is often used when it comes to income-related issues due to the impact of outliers (i.e. extreme values) on mean

One quantitative variable: measures of spread

- When we give a statistical summary of the values in a dataset, we are interested in not just the center of the data but also how spread out the data are.

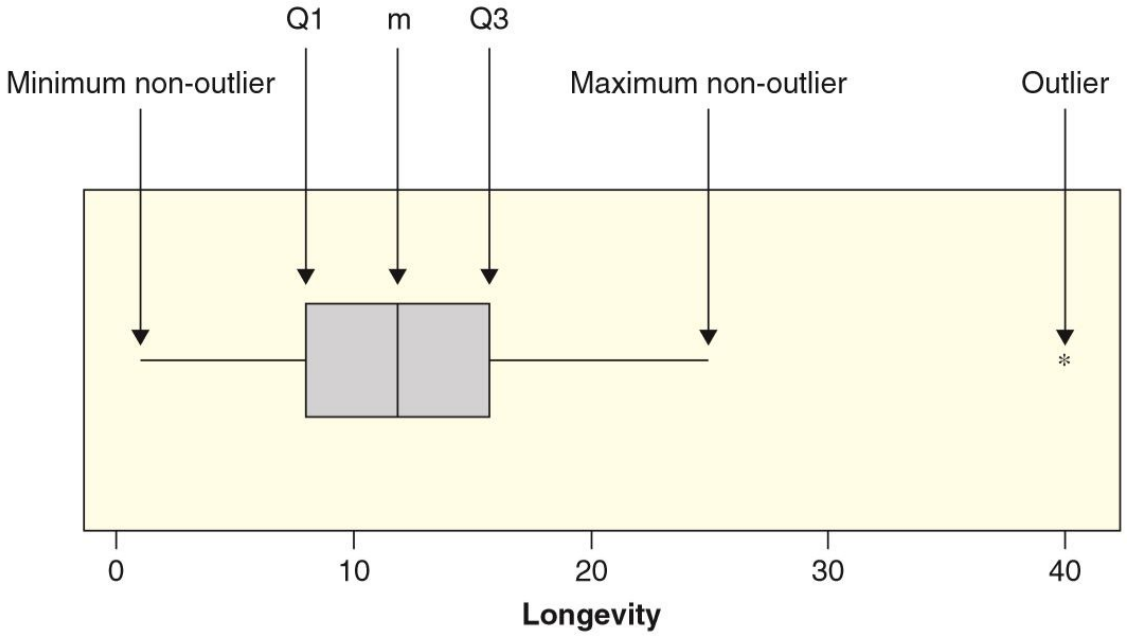


Q&A

Tukey's fences

for some nonnegative constant k. John Tukey proposed this test, where k=1.5 indicates an "outlier", and k=3 indicates data that is "far out"

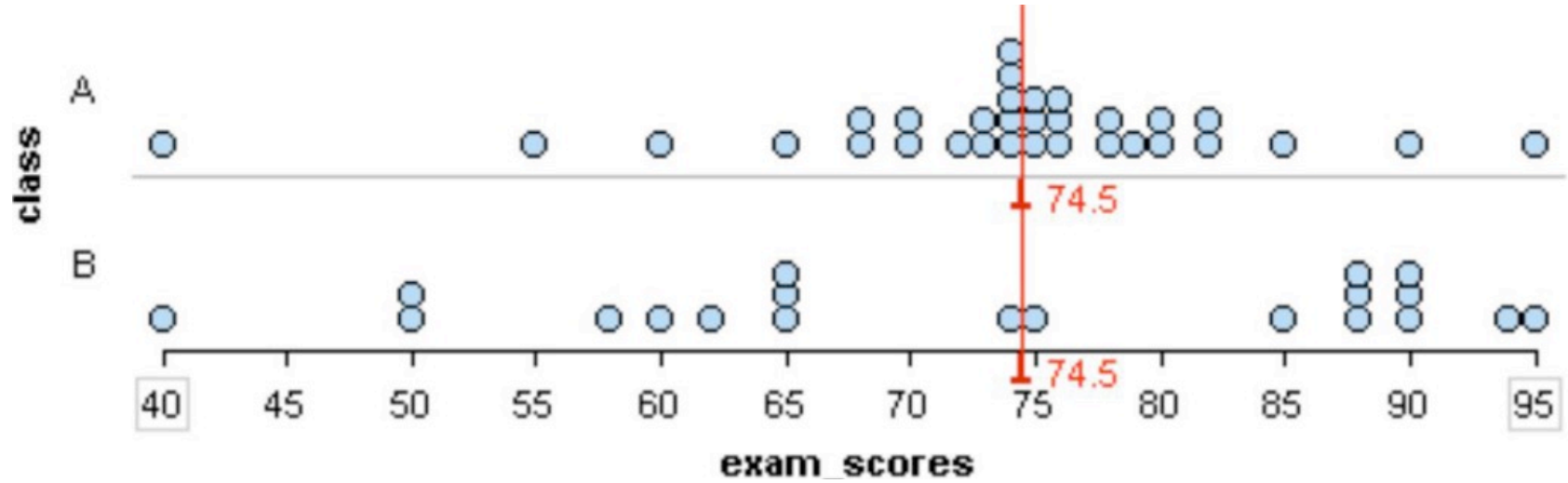
➢We compute Q1-1.5(IQR) = 8-1.5(8) = 8-12 =-4 and Q_3 + 1.5(*IQR*) = 16 + 1.5(8) = 16 + 12 = 28



1. *Tukey, John W (1977). Exploratory Data Analysis. Addison-Wesley. ISBN 978-0-201-07616-5. OCLC 3058187.*

Example -- Two sets of exam scores

- Consider the following two distributions of exam scores



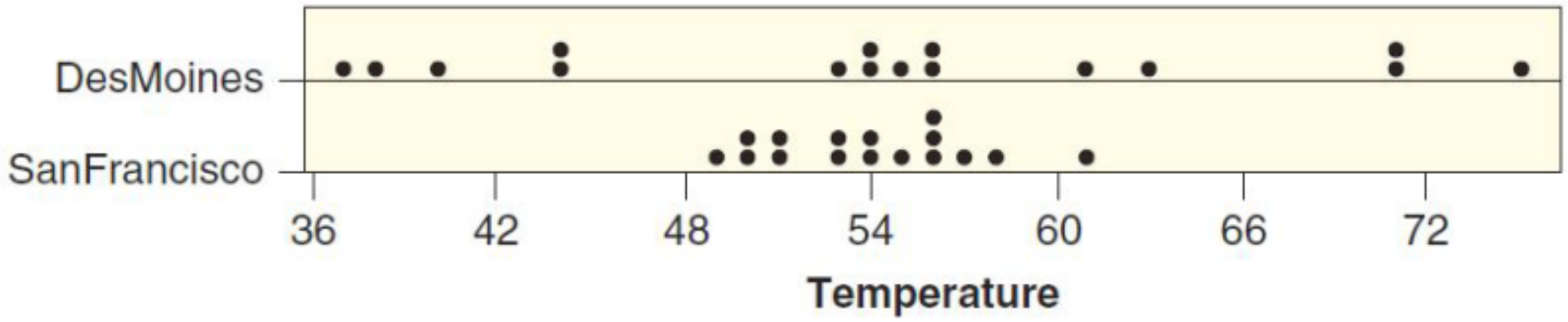
- Both distributions have a median of 74.5. Which distribution has more variability?
  - The answer to this question depends on how we measure variability.

Measures of spread

- For Des Moines the mean temperature is 54.49°F and the median is 54.50°F.
- For San Francisco the mean temperature is 54.01°F and the median is 54.0°F.

Des Moines	56.0	37.5	37.2	56.0	54.3	63.3	54.7	60.6
	70.6	53.7	52.9	74.9	44.4	40.3	44.4	71.0
San Francisco	51.0	55.3	55.7	48.7	56.2	57.2	49.5	61.0
	51.4	55.8	53.0	58.1	54.2	53.4	49.9	53.8

- The dotplots show that, while the centers may be similar, the distributions are very different.



Standard Deviation

The standard deviation for a quantitative variable measures the spread of the data in a sample.

- The standard deviation of a sample is denoted s, and measures how spread out the data are from the sample mean $\bar{x}$.

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

- The standard deviation of a population is denoted σ, which is the Greek letter "sigma" and measures how spread out the data are from the population mean μ.

$$\sigma = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n}}$$

Why the denominator is n-1 not n?

Biased estimation of σ

$$s = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n}}$$

Unbiased estimation of σ

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

- In statistics, **Bessel's correction** is the use of n − 1 instead of n in the formula for the sample variance and sample standard deviation, where n is the number of observations in a sample.
- This method corrects the bias in the estimation of the population variance. It also partially corrects the bias in the estimation of the population standard deviation.

$$(x_1 - \overline{x}, \ldots, x_n - \overline{x}),$$

One can understand Bessel‘s correction as the degrees of freedom in the residuals vector.
where $\bar{x}$ is the sample mean. While there are *n* independent observations in the sample, there are only *n* − 1 independent residuals, as they sum to 0.

Why n-1?

$$\begin{aligned}\sum_{i=1}^{n}(x_i-\overline{x})^2 &= \sum_{i=1}^{n}\left(x_i^2-2x_i\overline{x}+\overline{x}^2\right)\\ &= \sum_{i=1}^{n}x_i^2-2\overline{x}\sum_{i=1}^{n}x_i+\sum_{i=1}^{n}\overline{x}^2\\ &= \sum_{i=1}^{n}x_i^2-2n\overline{x}^2+n\overline{x}^2\\ &= \sum_{i=1}^{n}x_i^2-n\overline{x}^2\end{aligned}$$

so

$$\begin{aligned}\mathrm{E}\left(\sum_{i=1}^{n}\left[(x_i-\mu)-(\overline{x}-\mu)\right]^2\right) &= \mathrm{E}\left(\left(\sum_{i=1}^{n}(x_i-\mu)^2\right)-n(\overline{x}-\mu)^2\right)\\ &= \left(\sum_{i=1}^{n}\mathrm{E}\left((x_i-\mu)^2\right)\right)-n\,\mathrm{E}\left((\overline{x}-\mu)^2\right)\\ &= \left(\sum_{i=1}^{n}\mathrm{Var}(x_i)\right)-n\,\mathrm{Var}(\overline{x})\end{aligned}$$

$$\begin{aligned}\mathrm{E}(s^2) &= \mathrm{E}\left(\sum_{i=1}^{n}\frac{(x_i-\overline{x})^2}{n-1}\right)\\ &= \frac{1}{n-1}\,\mathrm{E}\left(\sum_{i=1}^{n}\left[(x_i-\mu)-(\overline{x}-\mu)\right]^2\right)\\ &= \frac{1}{n-1}\left[\left(\sum_{i=1}^{n}\mathrm{Var}(x_i)\right)-n\,\mathrm{Var}(\overline{x})\right]\end{aligned}$$

$$\mathrm{Var}(x_i)=\sigma^2$$

and also

$$\mathrm{Var}(\overline{x})=\frac{\sigma^2}{n}$$

$$\mathrm{E}(s^2)=\frac{1}{n-1}\left[\sum_{i=1}^{n}\sigma^2-n\sigma^2/n\right]=\frac{1}{n-1}(n\sigma^2-\sigma^2)=\sigma^2.$$

https://en.wikipedia.org/wiki/Bessel%27s_correction

How to calculate *s*

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

- Let us consider the simple data set: 2, 2, 4, 5, 6, 7, 9. We know that the mean is 5.

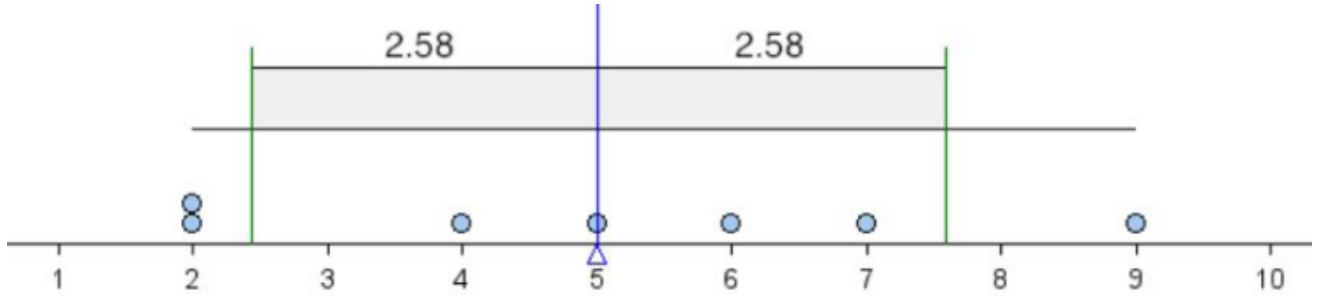
$2 - 5 = -3$
$2 - 5 = -3$
$4 - 5 = -1$
$5 - 5 = 0$
$6 - 5 = 1$
$7 - 5 = 2$
$9 - 5 = 4$

$(2 - 5)^2 = (-3)^2 = 9$
$(2 - 5)^2 = (-3)^2 = 9$
$(4 - 5)^2 = (-1)^2 = 1$
$(5 - 5)^2 = 0^2 = 0$
$(6 - 5)^2 = 1^2 = 1$
$(7 - 5)^2 = 2^2 = 4$
$(9 - 5)^2 = 4^2 = 16$

- Find the average of these squared differences.

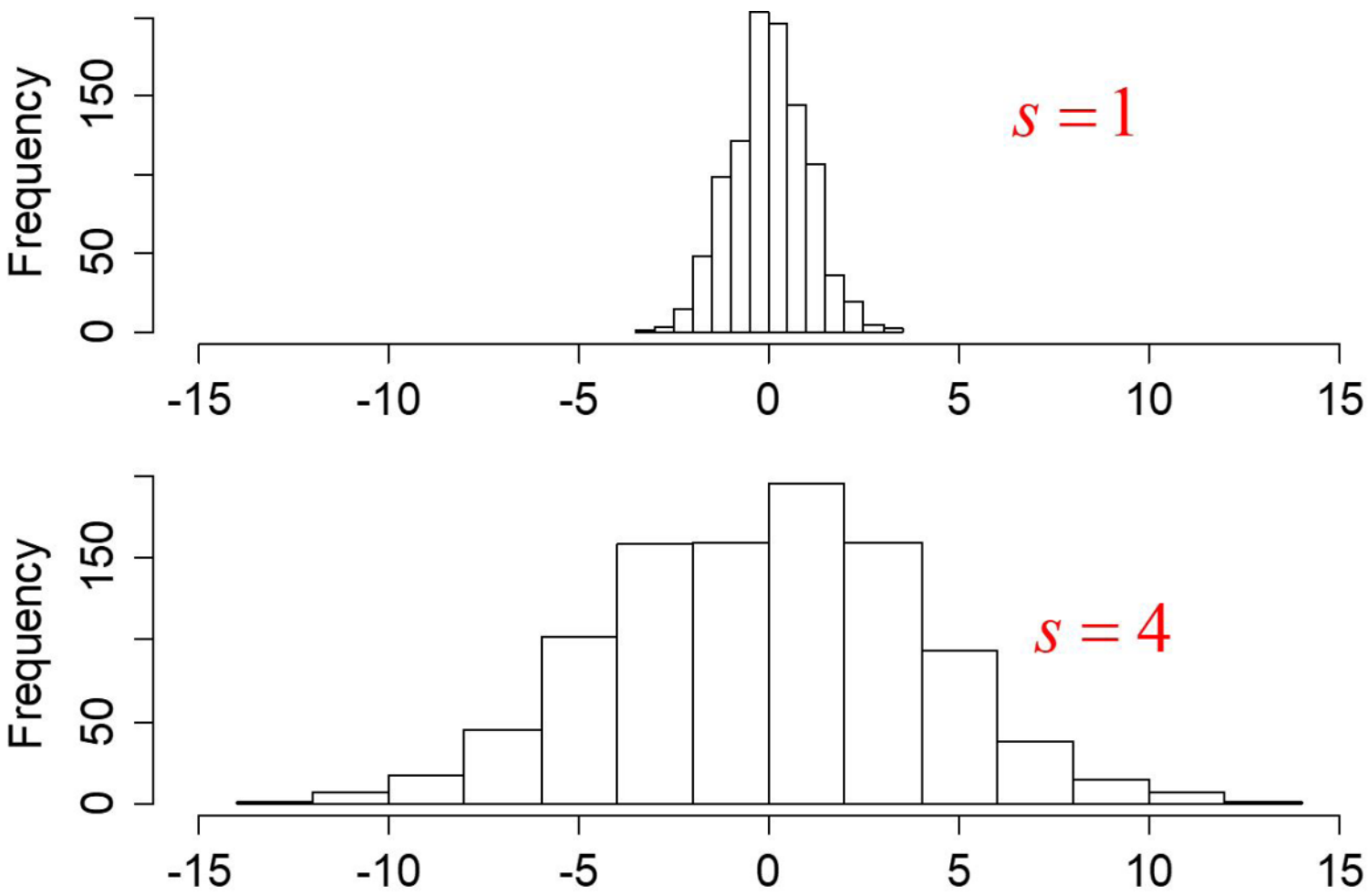
$$\frac{9 + 9 + 1 + 0 + 1 + 4 + 16}{6} \approx 6.67$$

- Take the square root of the value: $\sqrt{6.67} \approx 2.58$



Standard Deviation

- The larger the standard deviation, the more variability there is in the data and the more spread out the data are
- The standard deviation gives a rough estimate of the typical distance of a data values from the mean
- Both of these distributions are bell-shaped



Summary Statistics

- Measures of Center
  - Mean (not resistant)
  - Median (resistant)
- Measures of Spread
  - Standard deviation (not resistant)
  - IQR (resistant)
  - Range (not resistant)
- Most often, we use the mean and the standard deviation, because they are calculated based on all the data values, so use all the available information.

Summary: One Quantitative Variable

- Summary Statistics
  - Center: mean, median
  - Spread: standard deviation, range, IQR
  - Five number summary
- Visualization
  - Dotplot
  - Histogram
  - Boxplot
- Other concepts
  - Shape: symmetric, skewed, bell-shaped
  - Outliers, resistance