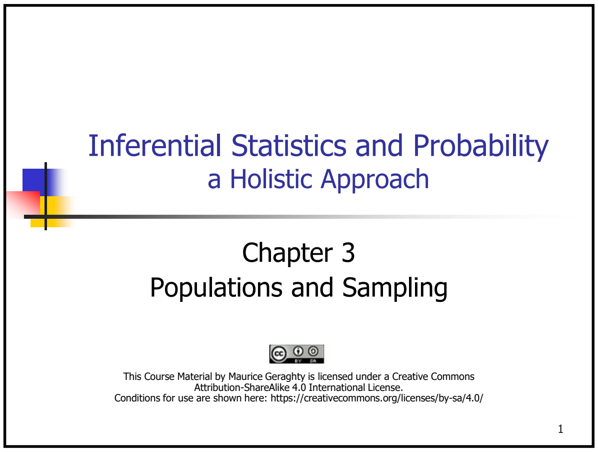
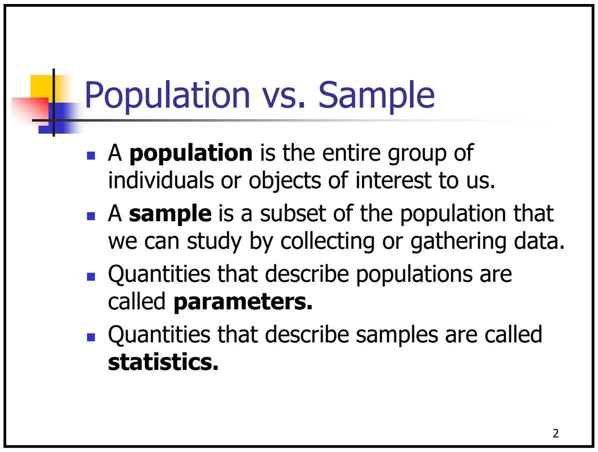


Inferential Statistics and Probability
a Holistic Approach

Chapter 3
Populations and Sampling

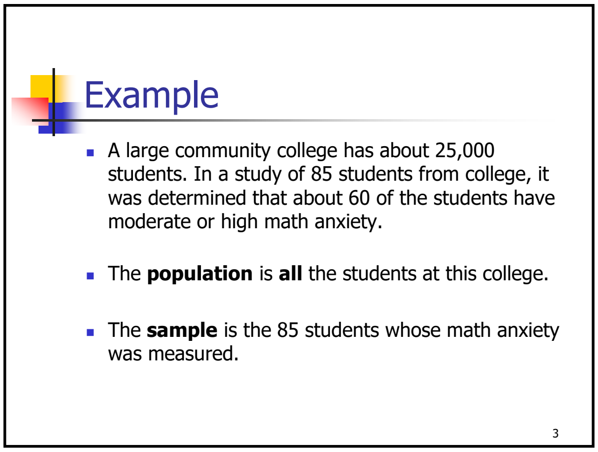
1



Population vs. Sample

- A **population** is the entire group of individuals or objects of interest to us.
- A **sample** is a subset of the population that we can study by collecting or gathering data.
- Quantities that describe populations are called **parameters.**
- Quantities that describe samples are called **statistics.**

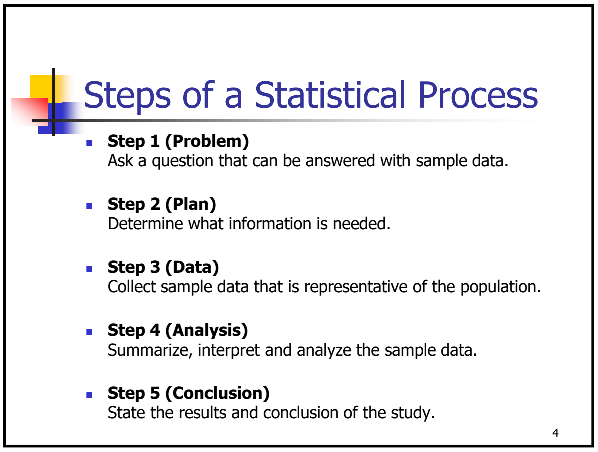
2



Example

- A large community college has about 25,000 students. In a study of 85 students from college, it was determined that about 60 of the students have moderate or high math anxiety.
- The **population** is **all** the students at this college.
- The **sample** is the 85 students whose math anxiety was measured.

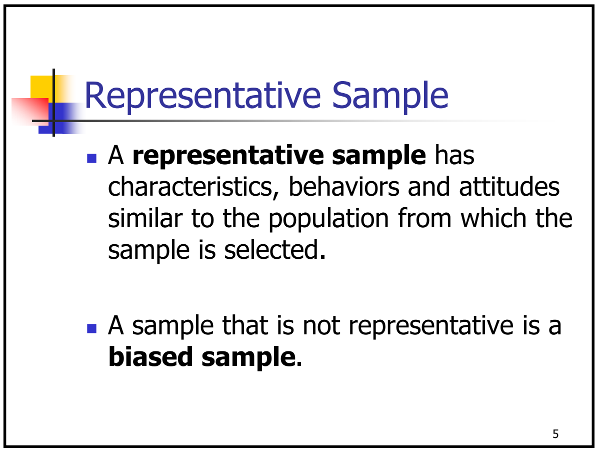
3



Steps of a Statistical Process

- **Step 1 (Problem)**
  Ask a question that can be answered with sample data.
- **Step 2 (Plan)**
  Determine what information is needed.
- **Step 3 (Data)**
  Collect sample data that is representative of the population.
- **Step 4 (Analysis)**
  Summarize, interpret and analyze the sample data.
- **Step 5 (Conclusion)**
  State the results and conclusion of the study.

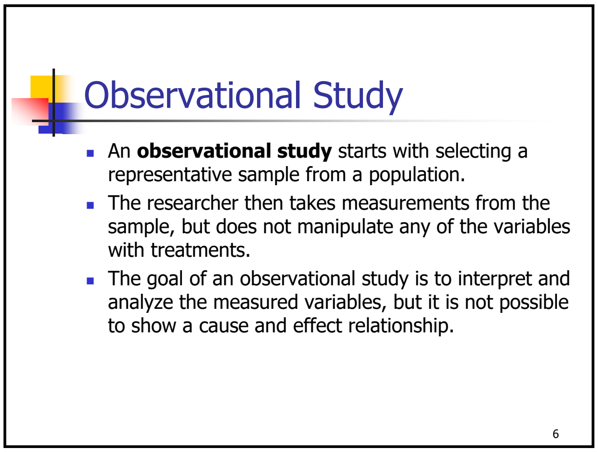
4



Representative Sample

- A **representative sample** has characteristics, behaviors and attitudes similar to the population from which the sample is selected.
- A sample that is not representative is a **biased sample**.

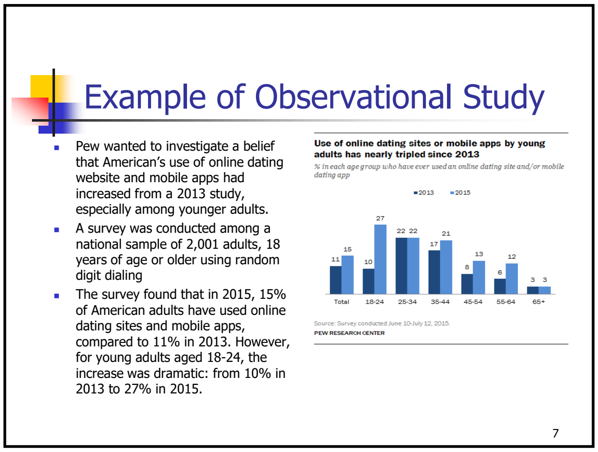
5



Observational Study

- An **observational study** starts with selecting a representative sample from a population.
- The researcher then takes measurements from the sample, but does not manipulate any of the variables with treatments.
- The goal of an observational study is to interpret and analyze the measured variables, but it is not possible to show a cause and effect relationship.

6



Example of Observational Study

- Pew wanted to investigate a belief that American's use of online dating website and mobile apps had increased from a 2013 study, especially among younger adults.
- A survey was conducted among a national sample of 2,001 adults, 18 years of age or older using random digit dialing
- The survey found that in 2015, 15% of American adults have used online dating sites and mobile apps, compared to 11% in 2013. However, for young adults aged 18-24, the increase was dramatic: from 10% in 2013 to 27% in 2015.

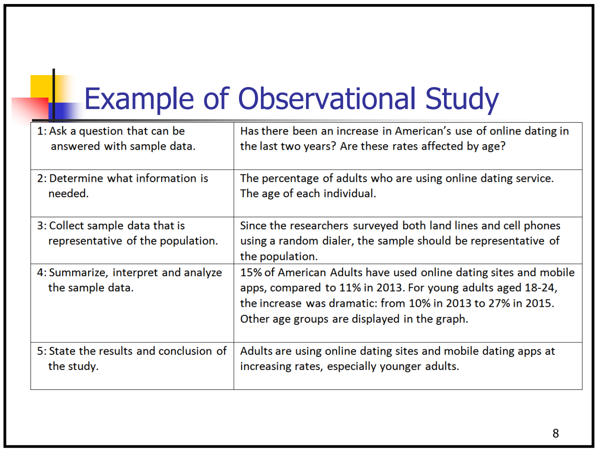
Use of online dating sites or mobile apps by young adults has nearly tripled since 2013

% in each age group who have ever used an online dating site and/or mobile dating app

Source: Survey conducted June 10-July 12, 2015.

PEW RESEARCH CENTER

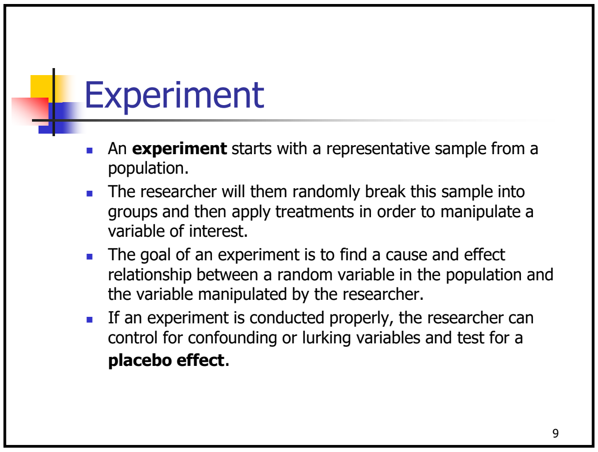
7



Example of Observational Study

1: Ask a question that can be answered with sample data.	Has there been an increase in American's use of online dating in the last two years? Are these rates affected by age?
2: Determine what information is needed.	The percentage of adults who are using online dating service. The age of each individual.
3: Collect sample data that is representative of the population.	Since the researchers surveyed both land lines and cell phones using a random dialer, the sample should be representative of the population.
4: Summarize, interpret and analyze the sample data.	15% of American Adults have used online dating sites and mobile apps, compared to 11% in 2013. For young adults aged 18-24, the increase was dramatic: from 10% in 2013 to 27% in 2015. Other age groups are displayed in the graph.
5: State the results and conclusion of the study.	Adults are using online dating sites and mobile dating apps at increasing rates, especially younger adults.

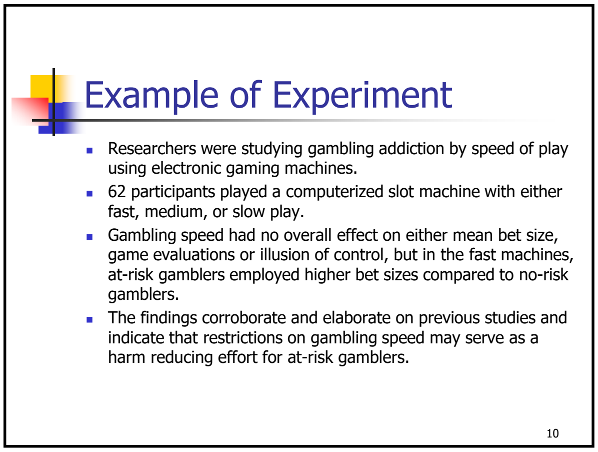
8



Experiment

- An **experiment** starts with a representative sample from a population.
- The researcher will them randomly break this sample into groups and then apply treatments in order to manipulate a variable of interest.
- The goal of an experiment is to find a cause and effect relationship between a random variable in the population and the variable manipulated by the researcher.
- If an experiment is conducted properly, the researcher can control for confounding or lurking variables and test for a **placebo effect.**

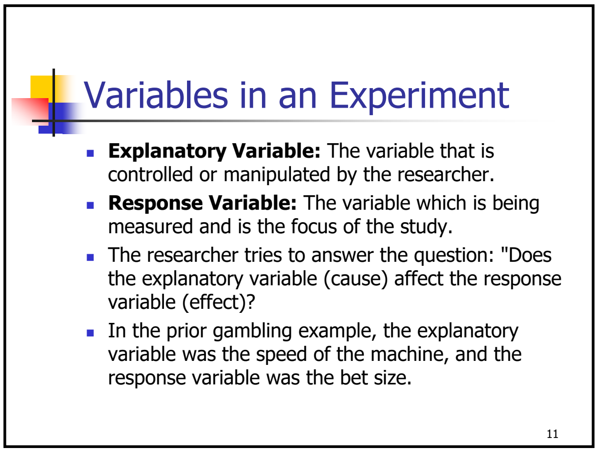
9



Example of Experiment

- Researchers were studying gambling addiction by speed of play using electronic gaming machines.
- 62 participants played a computerized slot machine with either fast, medium, or slow play.
- Gambling speed had no overall effect on either mean bet size, game evaluations or illusion of control, but in the fast machines, at-risk gamblers employed higher bet sizes compared to no-risk gamblers.
- The findings corroborate and elaborate on previous studies and indicate that restrictions on gambling speed may serve as a harm reducing effort for at-risk gamblers.

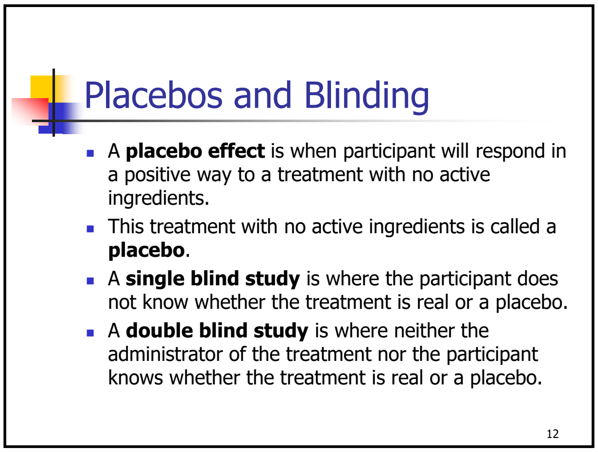
10



Variables in an Experiment

- **Explanatory Variable:** The variable that is controlled or manipulated by the researcher.
- **Response Variable:** The variable which is being measured and is the focus of the study.
- The researcher tries to answer the question: "Does the explanatory variable (cause) affect the response variable (effect)?
- In the prior gambling example, the explanatory variable was the speed of the machine, and the response variable was the bet size.

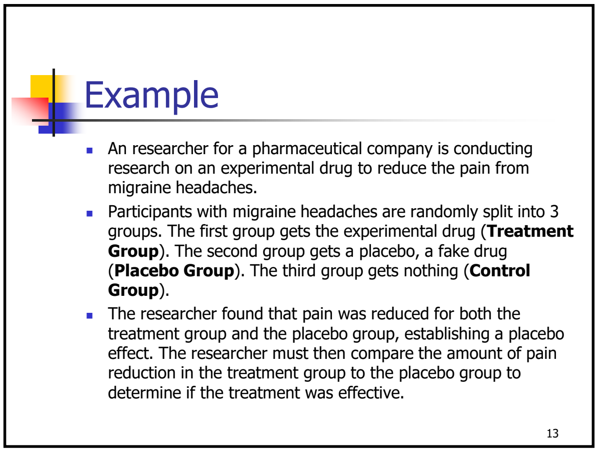
11



Placebos and Blinding

- A **placebo effect** is when participant will respond in a positive way to a treatment with no active ingredients.
- This treatment with no active ingredients is called a **placebo**.
- A **single blind study** is where the participant does not know whether the treatment is real or a placebo.
- A **double blind study** is where neither the administrator of the treatment nor the participant knows whether the treatment is real or a placebo.

12

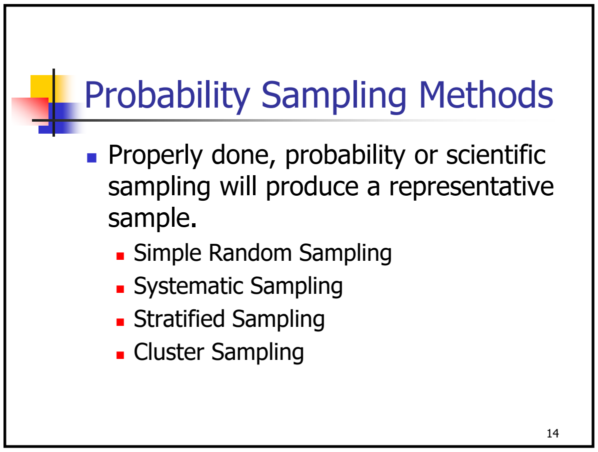


Example

- An researcher for a pharmaceutical company is conducting research on an experimental drug to reduce the pain from migraine headaches.
- Participants with migraine headaches are randomly split into 3 groups. The first group gets the experimental drug (**Treatment Group**). The second group gets a placebo, a fake drug (**Placebo Group**). The third group gets nothing (**Control Group**).
- The researcher found that pain was reduced for both the treatment group and the placebo group, establishing a placebo effect. The researcher must then compare the amount of pain reduction in the treatment group to the placebo group to determine if the treatment was effective.

13

13

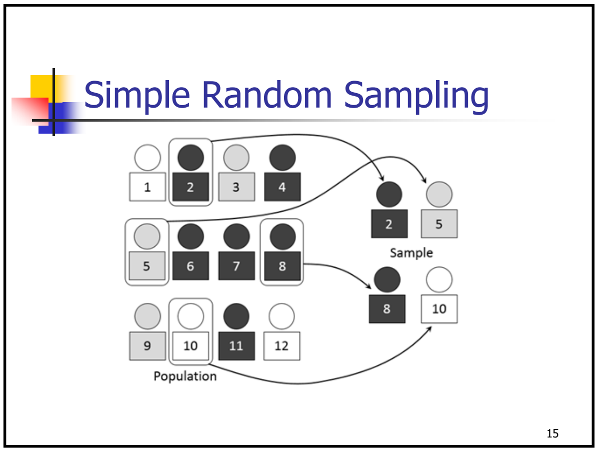


Probability Sampling Methods

- Properly done, probability or scientific sampling will produce a representative sample.
  - Simple Random Sampling
  - Systematic Sampling
  - Stratified Sampling
  - Cluster Sampling

14

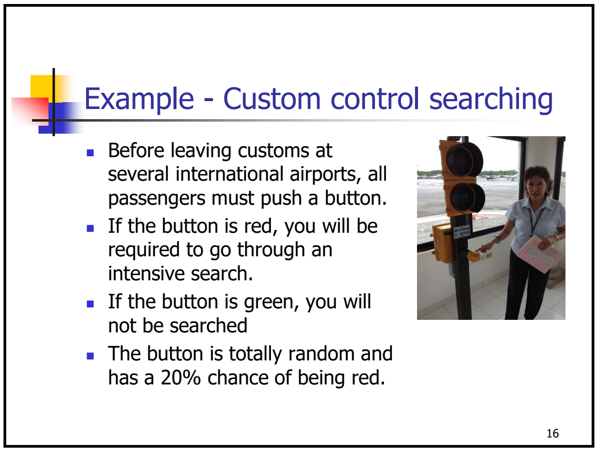
14



Simple Random Sampling

15

15

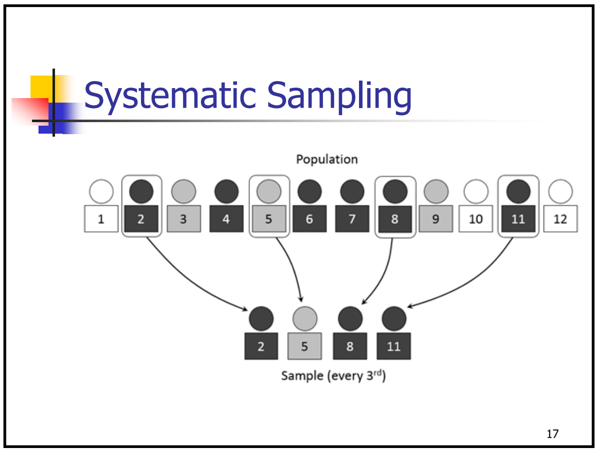


Example - Custom control searching

- Before leaving customs at several international airports, all passengers must push a button.
- If the button is red, you will be required to go through an intensive search.
- If the button is green, you will not be searched
- The button is totally random and has a 20% chance of being red.

16

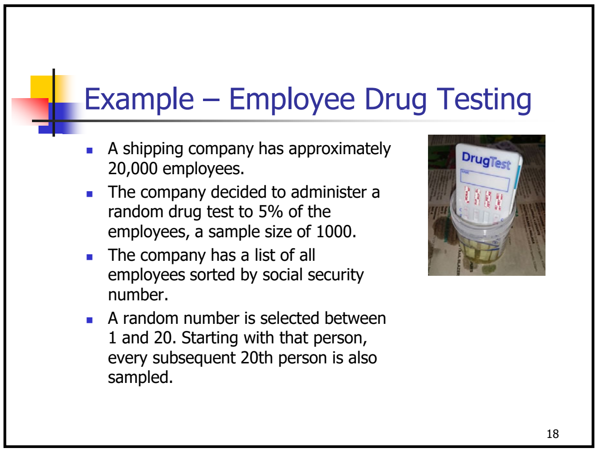
16



Systematic Sampling

17

17

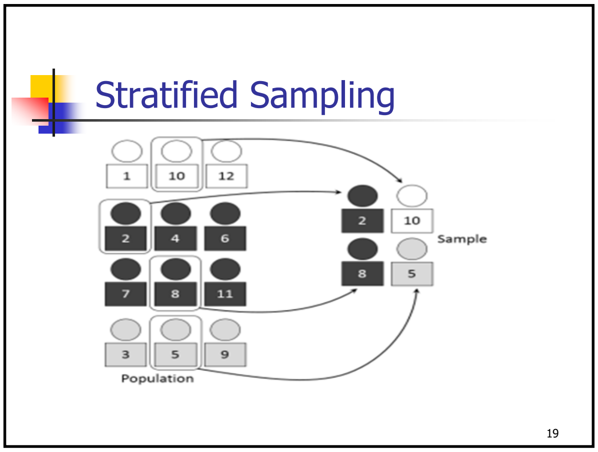


Example – Employee Drug Testing

- A shipping company has approximately 20,000 employees.
- The company decided to administer a random drug test to 5% of the employees, a sample size of 1000.
- The company has a list of all employees sorted by social security number.
- A random number is selected between 1 and 20. Starting with that person, every subsequent 20th person is also sampled.

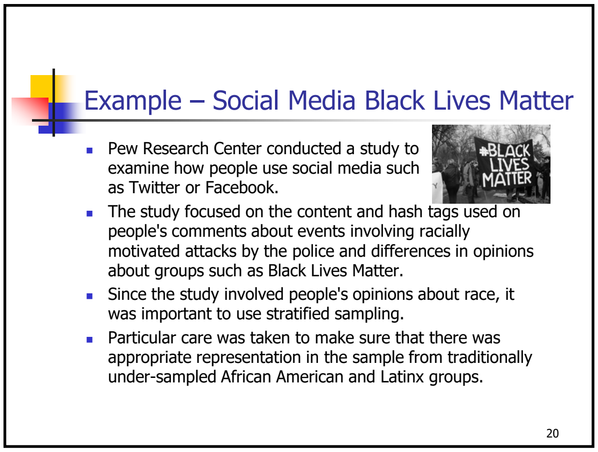
18

18



Stratified Sampling

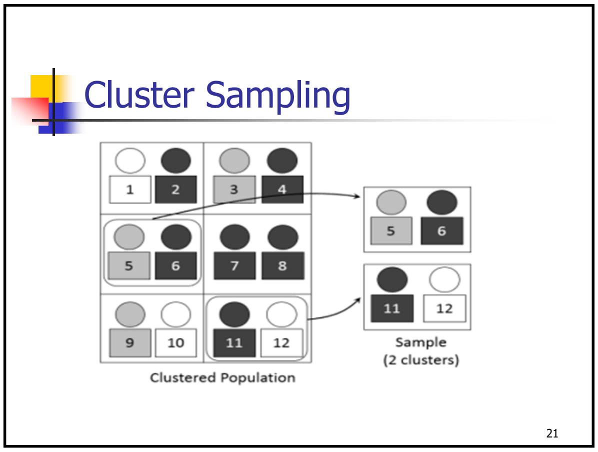
19



Example – Social Media Black Lives Matter

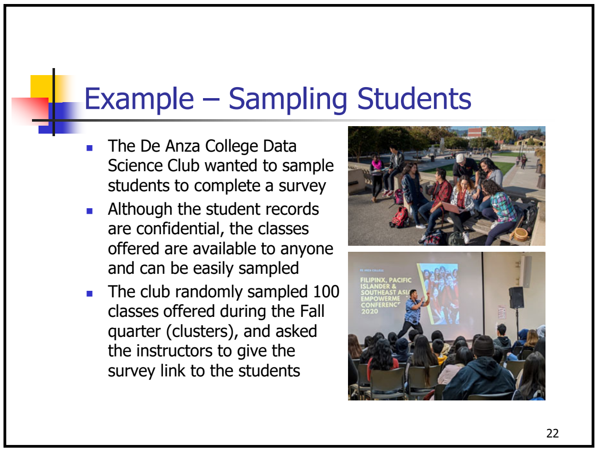
- Pew Research Center conducted a study to examine how people use social media such as Twitter or Facebook.
- The study focused on the content and hash tags used on people's comments about events involving racially motivated attacks by the police and differences in opinions about groups such as Black Lives Matter.
- Since the study involved people's opinions about race, it was important to use stratified sampling.
- Particular care was taken to make sure that there was appropriate representation in the sample from traditionally under-sampled African American and Latinx groups.

20



Cluster Sampling

21

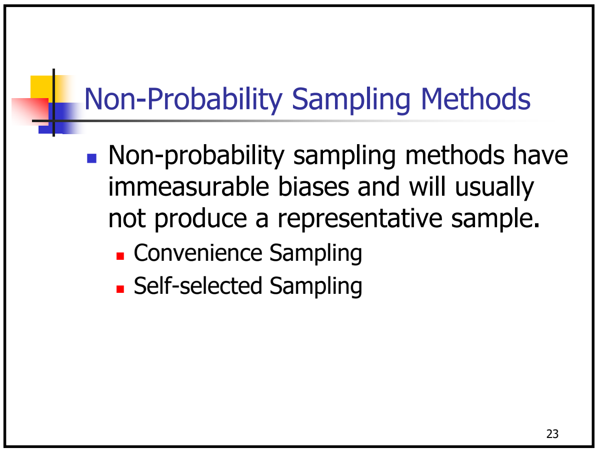


Example – Sampling Students

- The De Anza College Data Science Club wanted to sample students to complete a survey
- Although the student records are confidential, the classes offered are available to anyone and can be easily sampled
- The club randomly sampled 100 classes offered during the Fall quarter (clusters), and asked the instructors to give the survey link to the students

FILIPINX, PACIFIC ISLANDER & SOUTHEAST ASIAN EMPOWERMENT CONFERENCE 2020
22

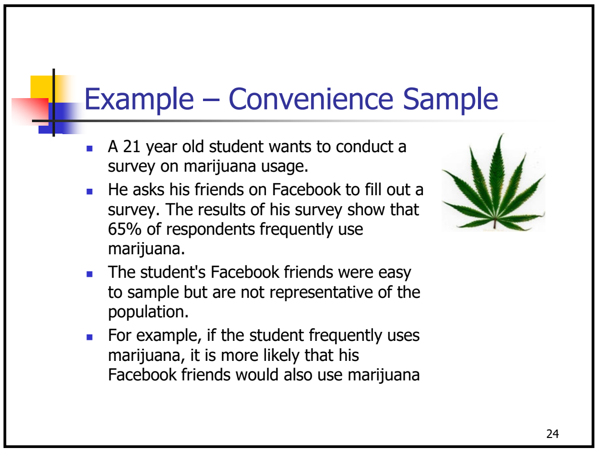
22



Non-Probability Sampling Methods

- Non-probability sampling methods have immeasurable biases and will usually not produce a representative sample.
  - Convenience Sampling
  - Self-selected Sampling

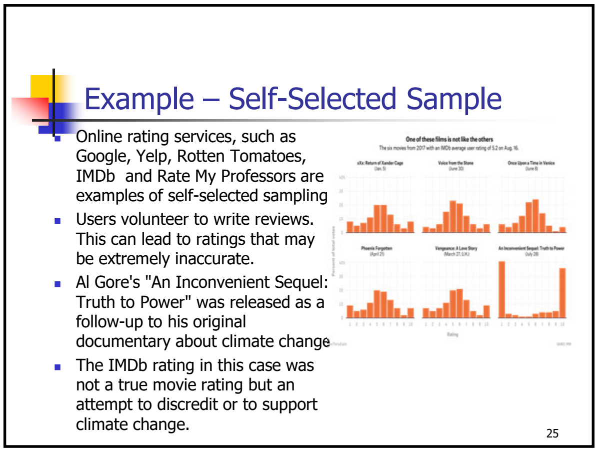
23



Example – Convenience Sample

- A 21 year old student wants to conduct a survey on marijuana usage.
- He asks his friends on Facebook to fill out a survey. The results of his survey show that 65% of respondents frequently use marijuana.
- The student's Facebook friends were easy to sample but are not representative of the population.
- For example, if the student frequently uses marijuana, it is more likely that his Facebook friends would also use marijuana

24



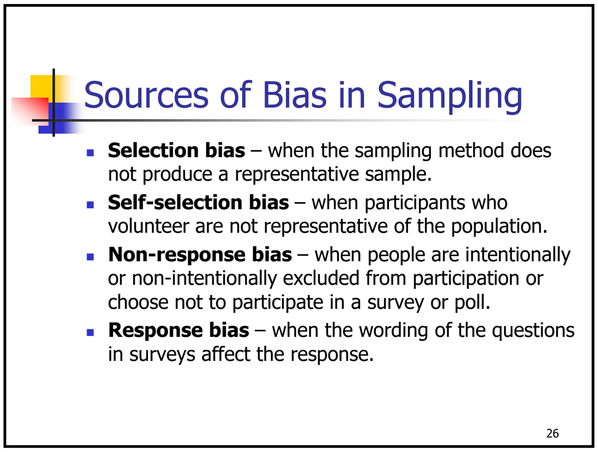
Example – Self-Selected Sample

- Online rating services, such as Google, Yelp, Rotten Tomatoes, IMDb and Rate My Professors are examples of self-selected sampling
- Users volunteer to write reviews. This can lead to ratings that may be extremely inaccurate.
- Al Gore's "An Inconvenient Sequel: Truth to Power" was released as a follow-up to his original documentary about climate change
- The IMDb rating in this case was not a true movie rating but an attempt to discredit or to support climate change.

One of these films is not like the others
The six movies from 2017 with an IMDb average user rating of 5.2 on Aug. 16.
xXx: Return of Xander Cage (Jan. 5)
Voice from the Stone (June 30)
Once Upon a Time in Venice (June 8)
Phoenix Forgotten (April 21)
Vengeance: A Love Story (March 27, U.K.)
An Inconvenient Sequel: Truth to Power (July 28)
Percent of total votes
Rating
1 2 3 4 5 6 7 8 9 10

25

25

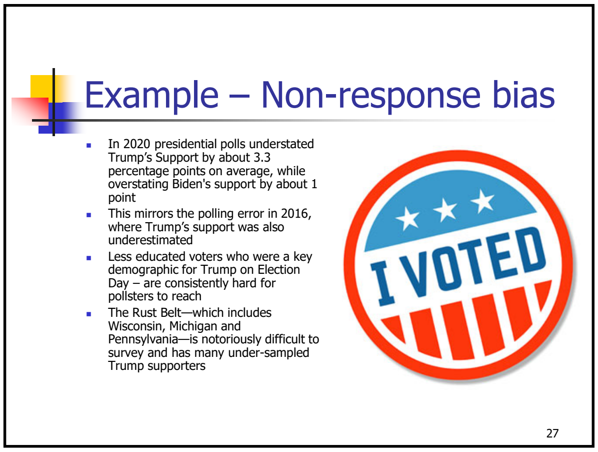


Sources of Bias in Sampling

- **Selection bias** – when the sampling method does not produce a representative sample.
- **Self-selection bias** – when participants who volunteer are not representative of the population.
- **Non-response bias** – when people are intentionally or non-intentionally excluded from participation or choose not to participate in a survey or poll.
- **Response bias** – when the wording of the questions in surveys affect the response.

26

26

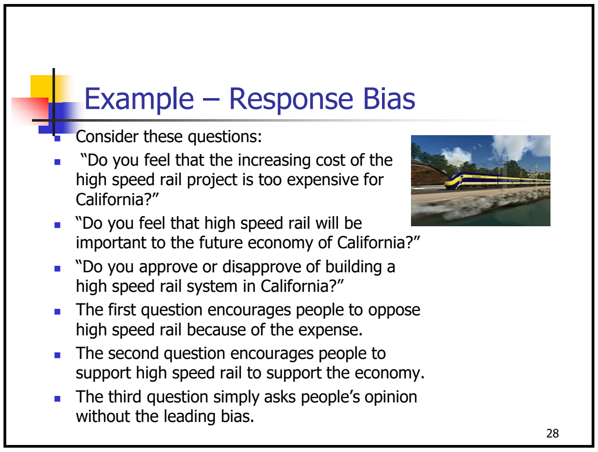


Example – Non-response bias

- In 2020 presidential polls understated Trump's Support by about 3.3 percentage points on average, while overstating Biden's support by about 1 point
- This mirrors the polling error in 2016, where Trump's support was also underestimated
- Less educated voters who were a key demographic for Trump on Election Day – are consistently hard for pollsters to reach
- The Rust Belt—which includes Wisconsin, Michigan and Pennsylvania—is notoriously difficult to survey and has many under-sampled Trump supporters

27

27



Example – Response Bias

- Consider these questions:
- "Do you feel that the increasing cost of the high speed rail project is too expensive for California?"
- "Do you feel that high speed rail will be important to the future economy of California?"
- "Do you approve or disapprove of building a high speed rail system in California?"
- The first question encourages people to oppose high speed rail because of the expense.
- The second question encourages people to support high speed rail to support the economy.
- The third question simply asks people's opinion without the leading bias.

28

28