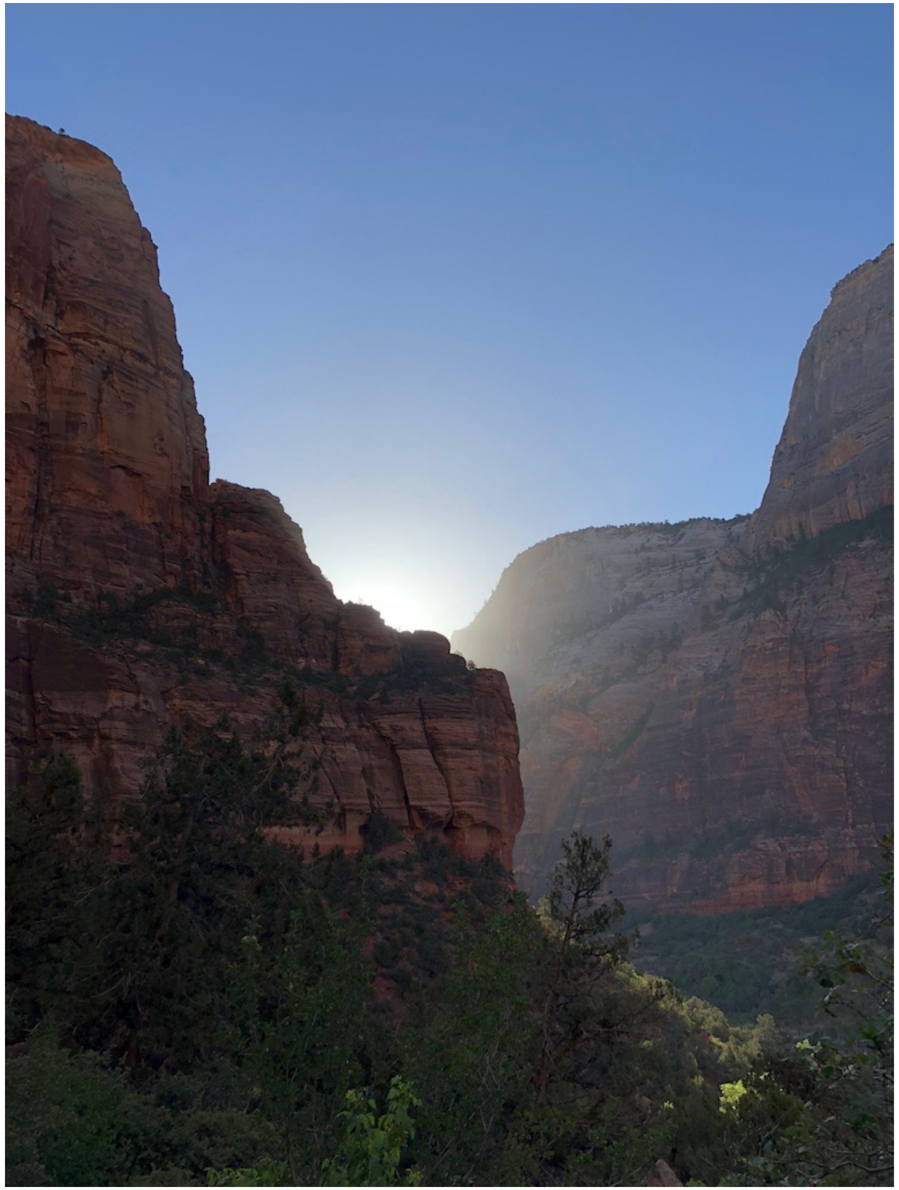
Yay!!! First Day of Stats Content!

Unit 1 – Basic Ideas

Your Excitedly Optimistic Professor Colton



Unit 1

Introduction

- Why Stats?

Introduction to Statistics

- Population vs. Sample
- Parameter vs. Statistic
- Descriptive vs Inferential

Methods of Data Collection

- Experimental vs Observational
- Data Collection Methods: Experiment, Simulation, Census, Sampling
- Sampling Methods: Simple Random, Cluster, Stratified, Systematic, Convenience, Volunteer Response
- Errors in Sampling (Sources of Bias)

Introduction – Why Statistics?

- Lets say we want to know if Ohio is a cat or dog state…
  - How can we figure this out?
  - Is that practical?
  - What do we do with all that info?
  - How can we make sense of it?

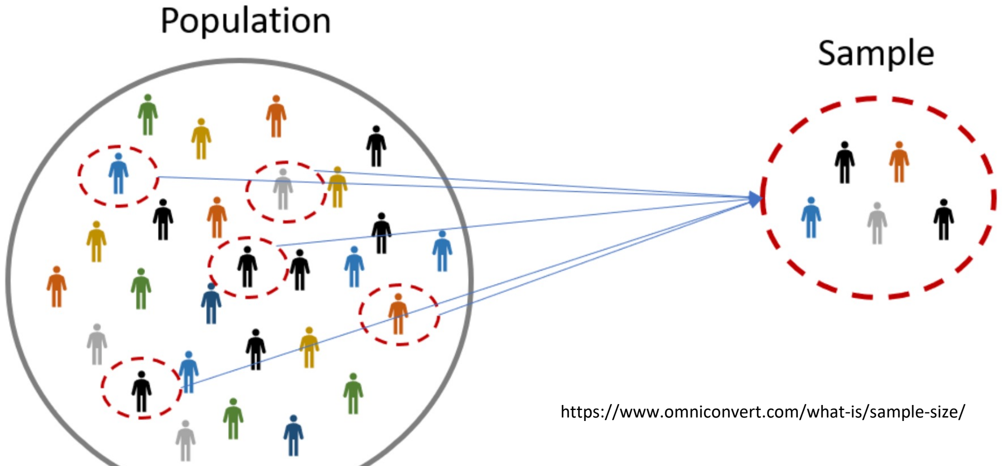
- Data! There's too much of it!
  - How do we learn from it?
  - Statistics!

Population vs Sample

- What is the difference between a Population and a Sample?

- **Population** is the set of all individuals/objects of interest
  - Ex) Back to Ohio cats vs dogs – Population is EVERY person in ALL of Ohio

- **A Sample** is a subset of individuals/objects from the population of interest
  - Ex) Everybody in ONLY Columbus



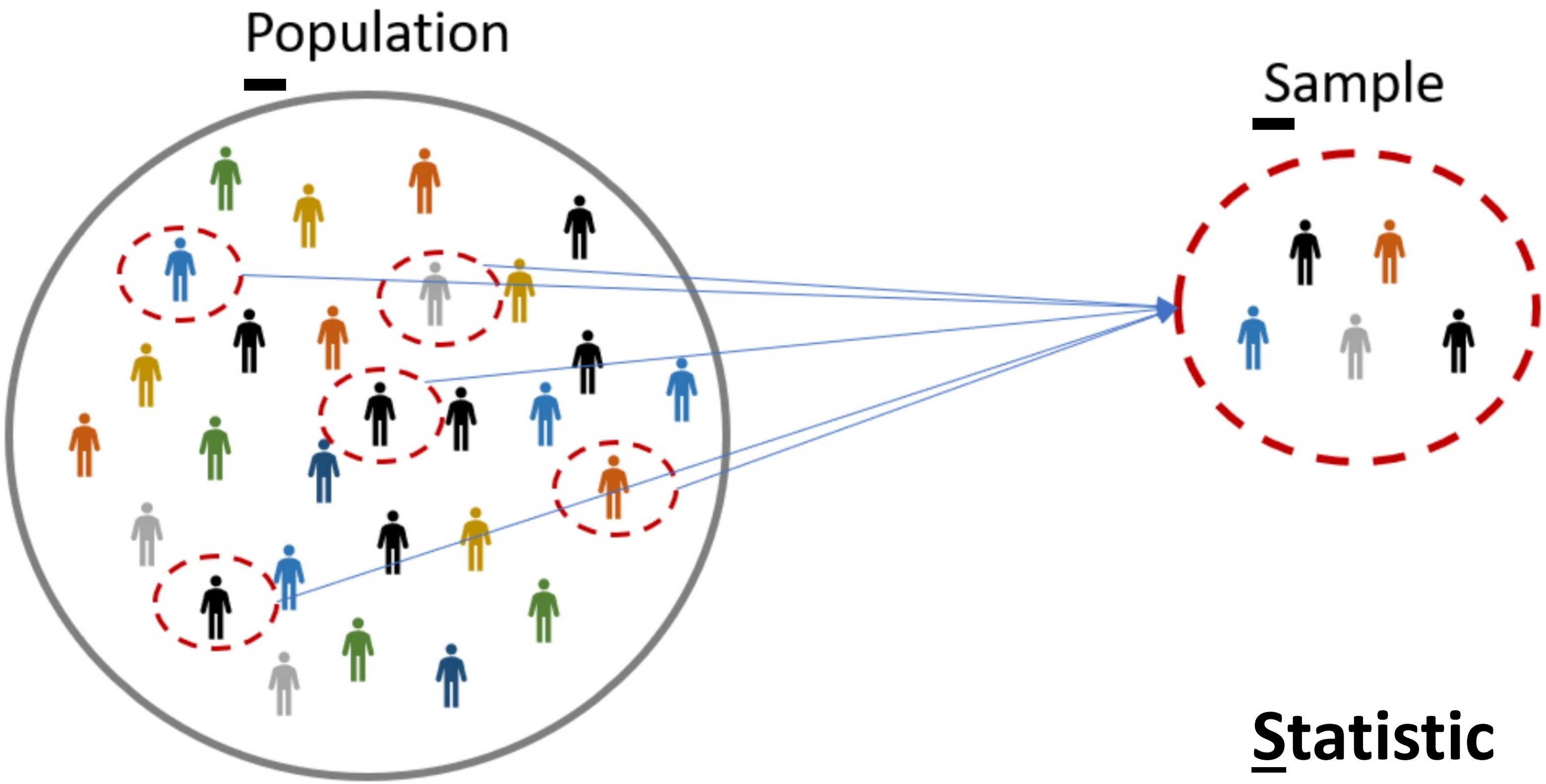
https://www.omniconvert.com/what-is/sample-size/

Parameter vs Statistic

- What is the difference between a Parameter and a Statistic?

- **Parameter** is a fixed numerical value that describes the Population
  - The OVERALL percentage of ALL people in Ohio that prefer cats (probably >> dogs ☺)
  - Why?? It describes the proportion for the ENTIRE population

- **A Statistic** is a numerical value that describes the sample that can vary
  - The percentage for people in Columbus or the percentage for Cincinnati
  - Would the percentages for these two cities be different?



Population

Sample

Statistic

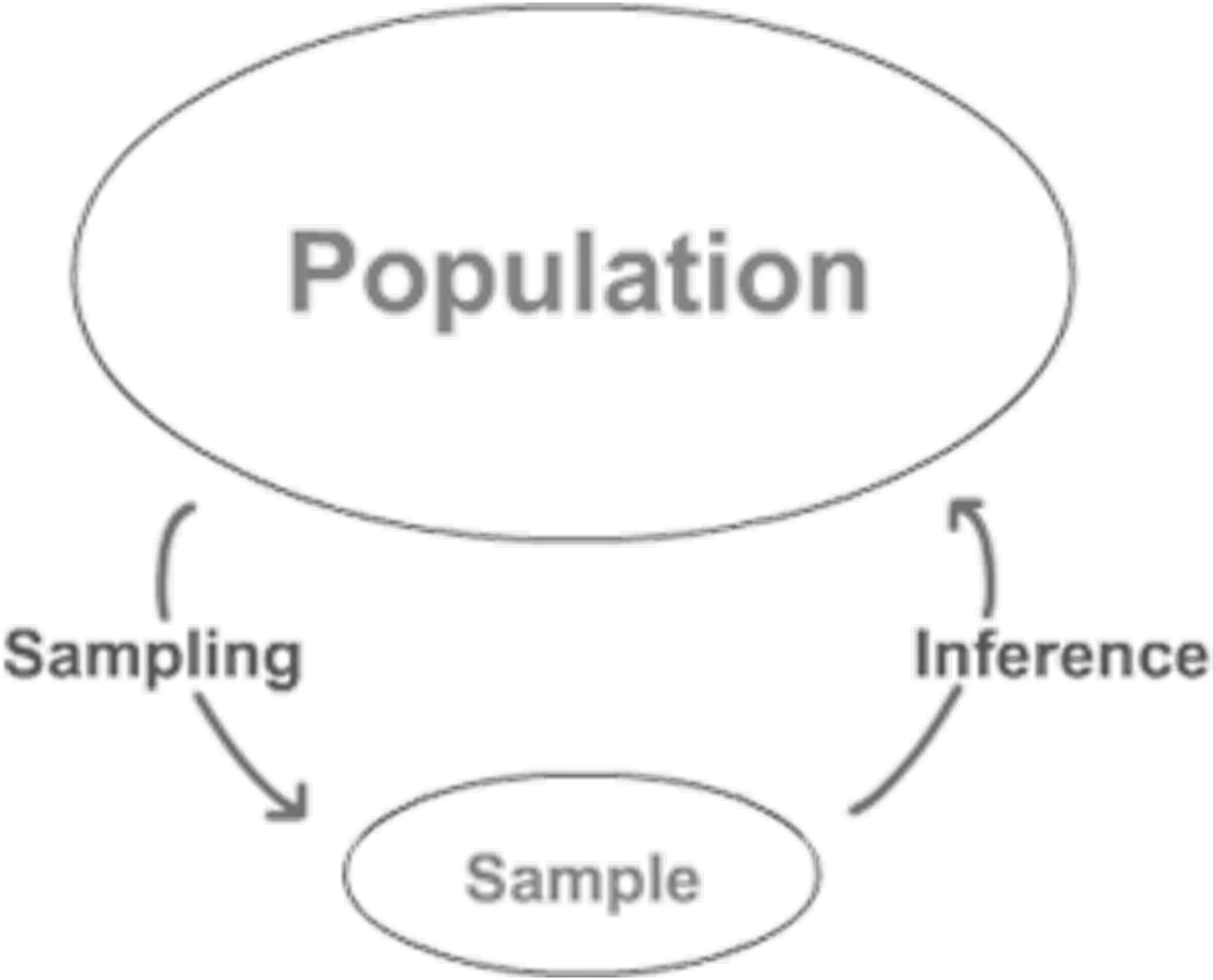
Parameter

https://www.omniconvert.com/what-is/sample-size/

Types of Studies by Goals

- Descriptive studies involve summarizing/describing the data
  - Mean, median, standard deviation, etc…
  - Graphical displays
  - Can be done on **BOTH** Sample and Population data
  - Ex: Drawing a bar graph of College Majors in STAT 1450

- Inferential studies involve drawing conclusions about a population based on a sample (infer something)
  - Confidence intervals
  - Hypothesis tests
  - Can **ONLY** be done with sample data
  - Ex: Is the true mean GPA of STAT 1450 students above 3.0?


Population
Sampling
Inference
Sample

Learning Check Quiz (LCQ): Descriptive vs Inferential

Descriptive or Inferential?

a) A researcher runs a controlled experiment of participants in a driving simulator. Half the participants are instructed to talk on their cell phone and the other half to talk to their passenger. The researcher records driving performance and wants to generalize the results to draw conclusions about distracted driving.

b) The Statistics department maintains records concerning all student evaluations of their instructors. A semester report includes graphical displays and summary statistics for the scores of the faculty members.

c) The Academic Integrity office keeps records on the occurrences of academic misconduct. Using past records, the office estimates how many incidents will occur during the upcoming school year.

Learning Check Quiz (LCQ): Descriptive vs Inferential

Descriptive or Inferential?

a) A researcher runs a controlled experiment of participants in a driving simulator. Half the participants are instructed to talk on their cell phone and the other half to talk to their passenger. The researcher records driving performance and wants to **generalize the results to draw conclusions about distracted driving.**
Inferential*, trying to draw a conclusion about driving performance.*

b) The Statistics department maintains records concerning **all** student evaluations of their instructors. A semester report includes **graphical displays and summary statistics** for the scores of the faculty members.
Descriptive*, showing graphs to describe student evals.*

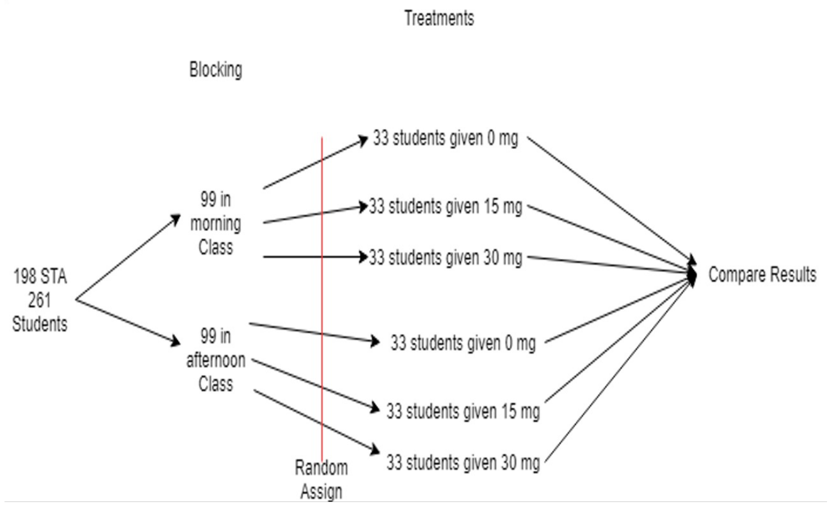
c) The Academic Integrity office keeps records on the occurrences of academic misconduct. Using past records, the office **estimates** how many incidents will occur during the upcoming school year.
Inferential*, using old results to predict next year (make a conclusion about next year's number of incidents)*

Types of Studies – By Data Collection Methods

Experimental Studies

- ***Imposes*** treatments and then observes results.
- Can help determine a causation
- Reduces the potential a lurking variable can affect the results (controls for them)
- More accurate to determine a relationship between the explanatory variable and response.

- Ex) Clinical trials



Observational Studies

- **Observe** what happens without imposing restraints.
- No random assignment
- Can reveal an association or correlation between variables but not causation.
    - Causation requires a lot of extra work and research

- Ex) Car accidents at an intersection

Types of Studies – By Data Collection Methods

Experiments

- ***Imposes*** treatments and records response

Simulation

- Not always practical to do an experiment (cost, time, etc.)
- Simulations allows us to use computers to mimic what would happen in real life, or what we believe happens in real life.
    - The Matrix
- Can use computers to simulate or use “devices” that are random
    - Dice
    - Coin
    - Deck of Cards

Census

- Collect data from EVERY member of the population
- Observational study on the ENTIRE population
- Ex) US Census

Sampling

- Collect data from a SUBSET of the population of interest
- Observational study on a SAMPLE (a PORTION of the population
- Lots of different ways to do this!
- Ex) Randomly selecting 30 CSCC students

Sampling Methods

Sample Surveys

- **Sample surveys** are an important kind of observational study.
- They survey some group of individuals by studying only some of its members.
- Individuals are selected not because they are of special interest, but because they represent the larger group (the population).

Good ones!

- All statistical sampling designs use chance, rather than human choice, to select the sample

- Simple Random Samples (SRS)
- Stratified Sampling
- Cluster Sampling
- Systematic Sampling

Bad ones ☹

- Convenience Sampling
- Volunteer Response Sampling

Sampling Methods

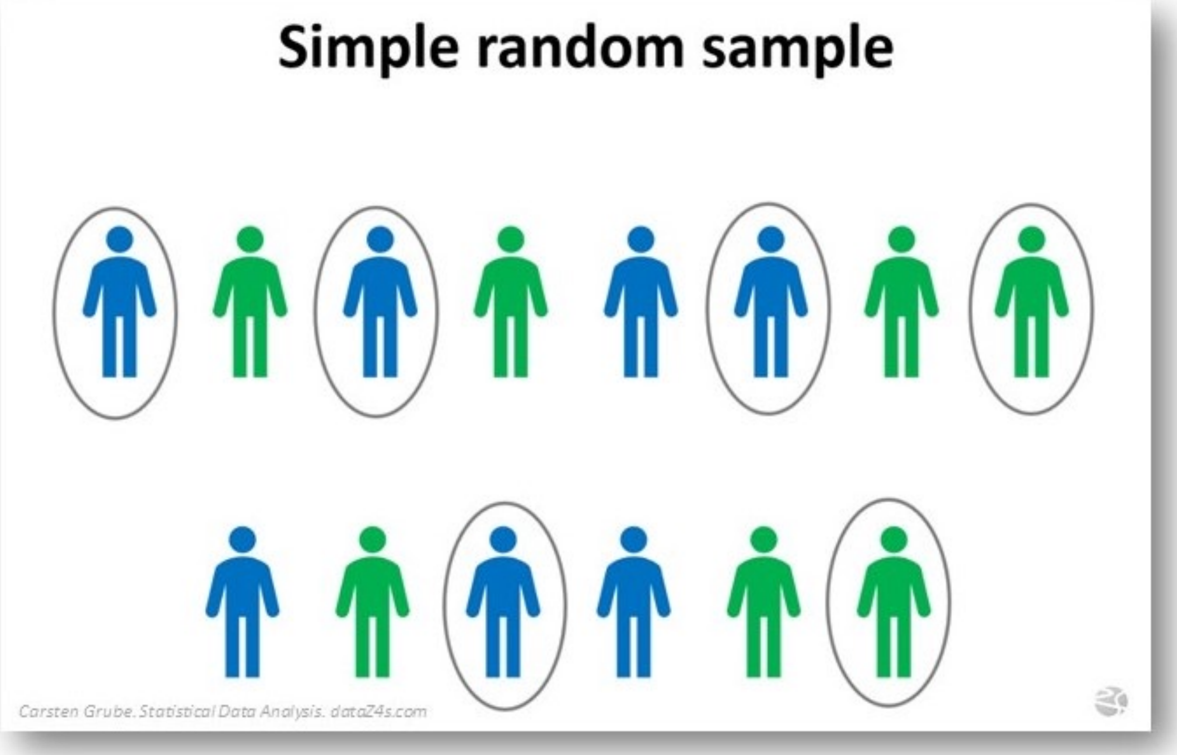
Simple random samples (SRS)

Two key parts to taking an SRS:

1. Every individual in the population has an equal chance of being selected
2. Every possible sample of the size we plan to draw has an equal chance to be selected
   - In other words, each combination of people has an equal chance of being selected

- Sampling frame – a list of individuals from which the sample is drawn
  - Assign a random number to each individual in the sampling frame

- Ex) Use a random number generator to select names from a list

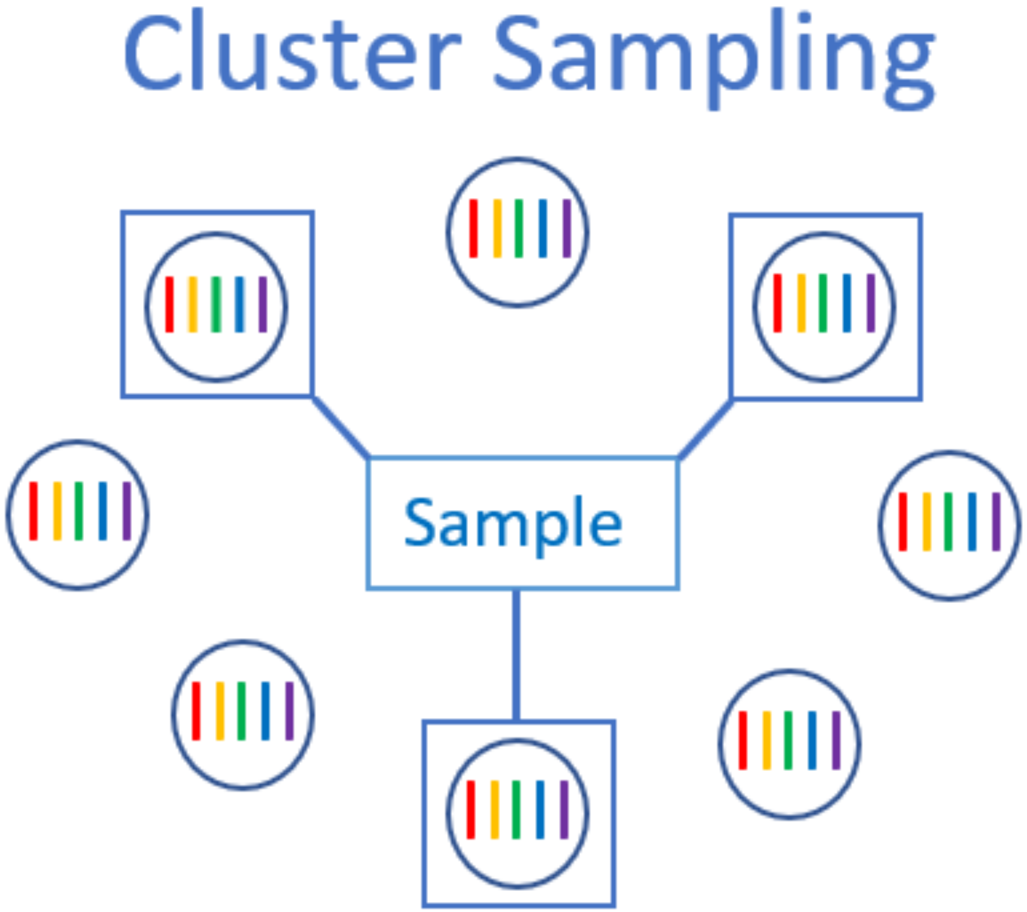


More Sampling Methods

Cluster Sampling

1. Split the population into **representative groups** called clusters
2. Use random sampling to select several clusters
3. Perform a census of each selected cluster

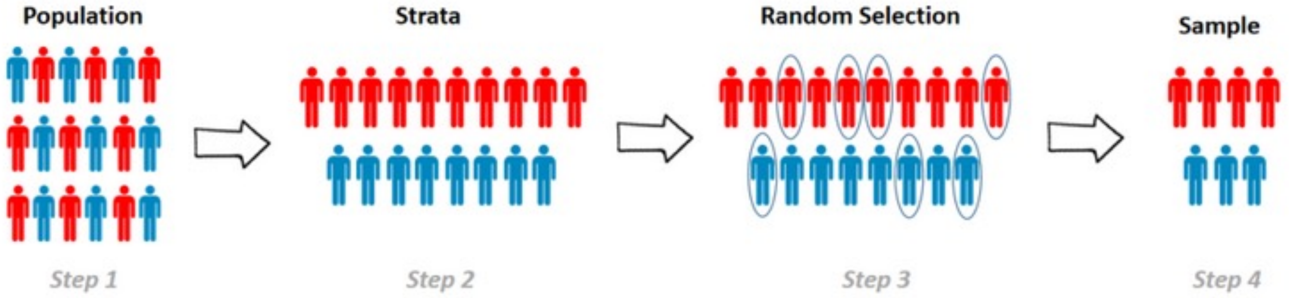
- If each cluster represents the full population fairly, cluster sampling will give an unbiased sample
- **Clusters are heterogeneous and resemble the overall population**
- Cluster sampling is usually more practical and affordable



https://statisticsbyjim.com/basics/cluster-sampling/

More Sampling Methods

Stratified Random Sampling



- Strata – homogeneous groups
  - Used to reduce sampling variability
  - Ensures that the proportions of characteristic(s) in our sample match the proportions of the characteristic(s) in the population
  - Helps us to see differences among groups
  - Sampling procedure may be more difficult

1. Stratify the population – Divide the population into homogeneous groups
2. Use SRS to choose members from each strata
3. Combine the groups from each strata to form your sample
   - Each stratum is different from the others

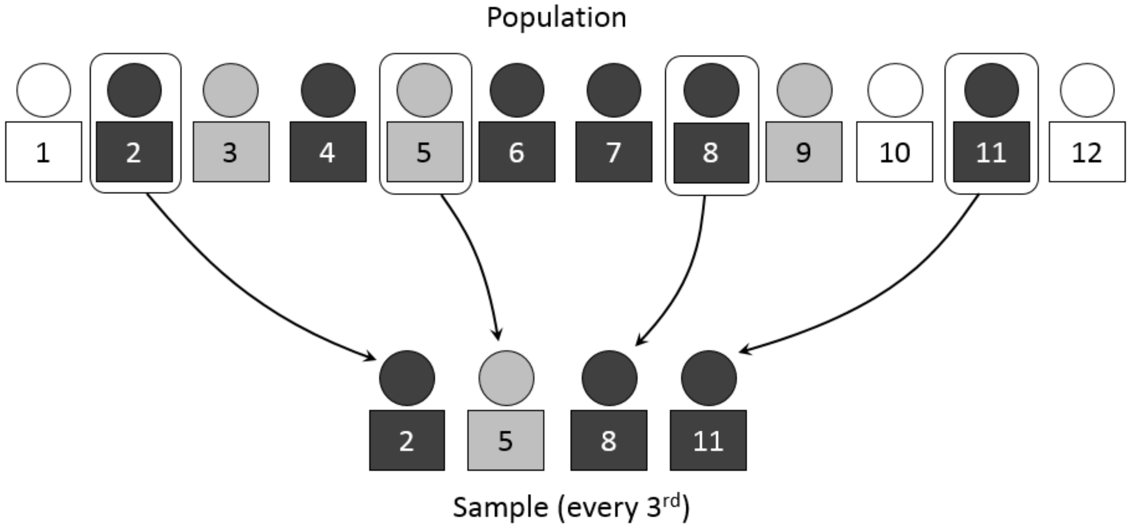
https://uedufy.com/what-is-stratified-sampling/

More Sampling Methods

Systematic Sampling

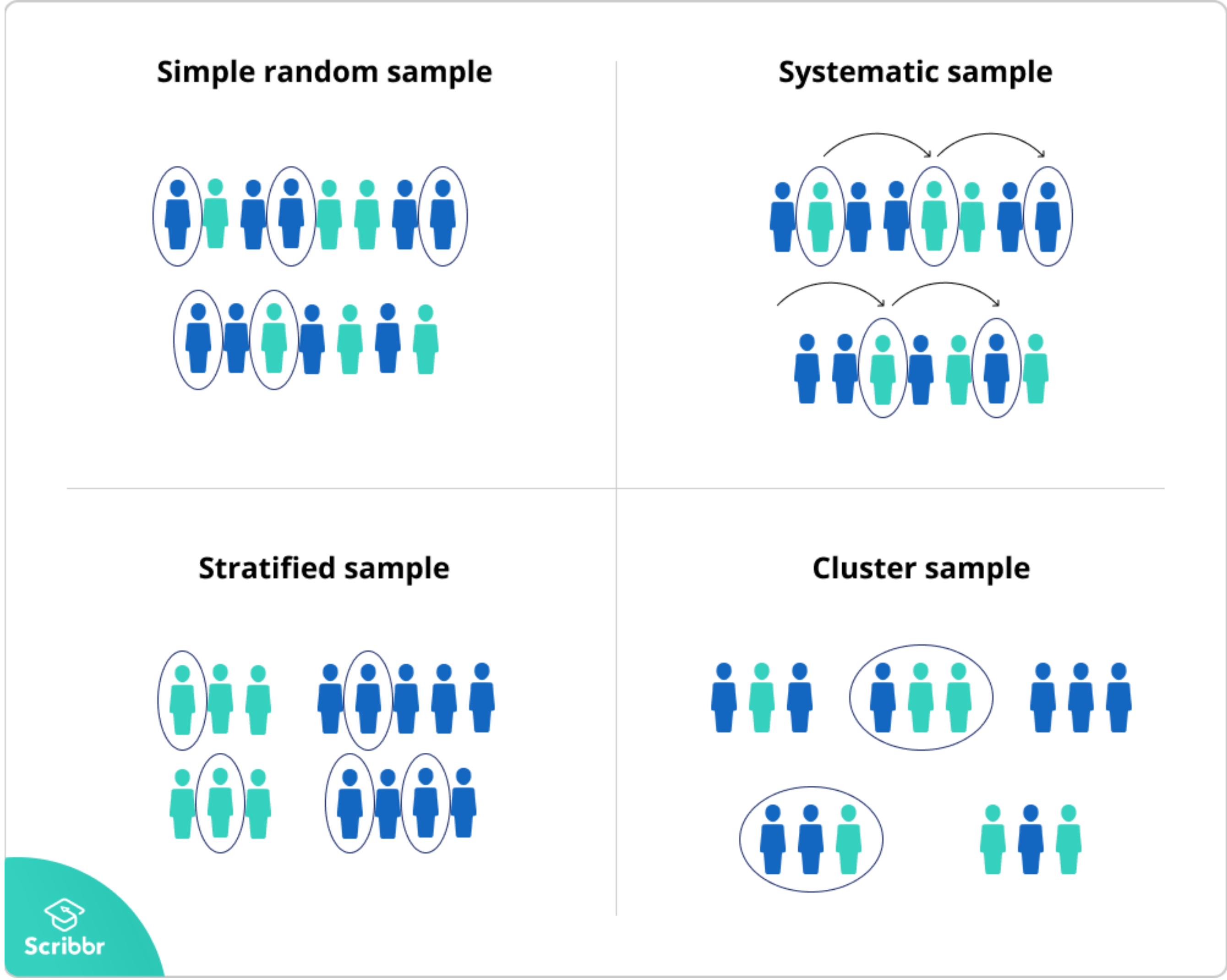
- Much less expensive than SRS
- Must justify the assumption that the systematic method is NOT associated with any of the measured variables

1. Randomly select a starting place
2. Employ a systematic method to continue choosing your sample
   - For example, take every 20th name on a list of names
   - The order of the list cannot be associated in any way with the responses sought
   - Beware of confounding variables



https://www.netquest.com/blog/en/systematic-sampling

Summary of Sampling Methods



https://anthonybmasters.medium.com/systematic-sampling-c65bb51e29dd

Sampling Methods Example

Example

I wish to determine the proportion of the STAT 1450 class that has a Mac laptop. How can I go about getting data to answer this question from our class?

- Simple Random Sample
  - Randomly ask 10 students from class what laptop they use.
- Systematic Sampling
  - Ask every 4th person to that walks into class what type of laptop they use.
- Cluster Random Sampling
  - Randomly sample 3 breakout groups; ask everyone in that (take a census of the) group about what type of laptop they use.
- Stratified Random Sampling
  - Randomly sample 5 students from the list of Freshman and Sophomores respectively and ask them what type of laptop they use.

LCQ: Sampling Methods

You are tasked with conducting a survey to answer the question, “What is the favorite subject of students who attend East High School?” Describe how you could obtain a sample to answer this question using each of the following types of sampling methods listed below.

Simple Random Sample:

Systematic Sample:

Cluster Sample:

Stratified Random Sample:

LCQ: Sampling Methods

You are tasked with conducting a survey to answer the question, “What is the favorite subject of students who attend East High School?” Describe how you could obtain a sample to answer this question using each of the following types of sampling methods listed below.

Possible answers:

Simple Random Sample:

- *Randomly select 30 students from the entire student body.*
- *Assign random numbers to all students, use random number generator and 20 pick numbers*

Systematic Sample:

- *Ask every 10th student that arrives that morning*
- *Assign numbers to all students, ask every fifth student*

Cluster Sample:

- *Randomly select 5 classrooms and ask every student in each of those classrooms*
- *Randomly select a few lunch periods to and ask all students there → this works because we can assume that there is a mix of students from all grades in each lunch period)*
- *Randomly select 5 buses and survey all students on the bus → again it is reasonable to believe that there are lots of students of different types on each bus, so each bus is like a mini population that is ~ representative of the entire school*

Stratified Random Sample:

- *Divide the students based on class (Freshman, Sophomore, …), then randomly sample 10 students from each class*
- *Divide students based on Male / Female, then randomly sample within each group → this is good because students within each group are similar (homogeneous), and we then random sample*

Harder LCQ: Sampling Methods

A researcher wants to study regional differences in dental care. He takes a multistage sample by dividing the United States into four regions, taking a simple random sample of ten schools in each region, randomly sampling three classrooms in each school, and interviewing all students in those classrooms. Identify the type of sampling employed in each stage of this sampling design. (Agresti & Franklin, 2013, p. 192)

WEST
MIDWEST
NORTHEAST
Seattle-Puget Sound
PACIFIC
MOUNTAIN
WEST NORTH CENTRAL
EAST NORTH CENTRAL
Detroit
NEW ENGLAND
San Francisco-Oakland
San Jose-Monterey
Los Angeles
SOUTH ATLANTIC
EAST SOUTH CENTRAL
Atlanta
WEST SOUTH CENTRAL
Division
State
SOUTH

Stage(s)???:

1)

2)

3)

4)

5)

Harder LCQ: Sampling Methods

A researcher wants to study regional differences in dental care. He takes a multistage sample by dividing the United States into four regions, taking a simple random sample of ten schools in each region, randomly sampling three classrooms in each school, and interviewing all students in those classrooms. Identify the type of sampling employed in each stage of this sampling design. (Agresti & Franklin, 2013, p. 192)

WEST
MIDWEST
NORTHEAST
NEW ENGLAND
Seattle-Puget Sound
PACIFIC
MOUNTAIN
WEST NORTH CENTRAL
Detroit
EAST NORTH CENTRAL
San Francisco-Oakland
San Jose-Monterey
Los Angeles
SOUTH ATLANTIC
EAST SOUTH CENTRAL
Atlanta
WEST SOUTH CENTRAL
Division
State
SOUTH

Stage(s):

1) Stratified → this is stratified (not cluster) because we think there are regional differences, so states within each region are similar in terms of dental care. And the regions are our different strata

2) SRS

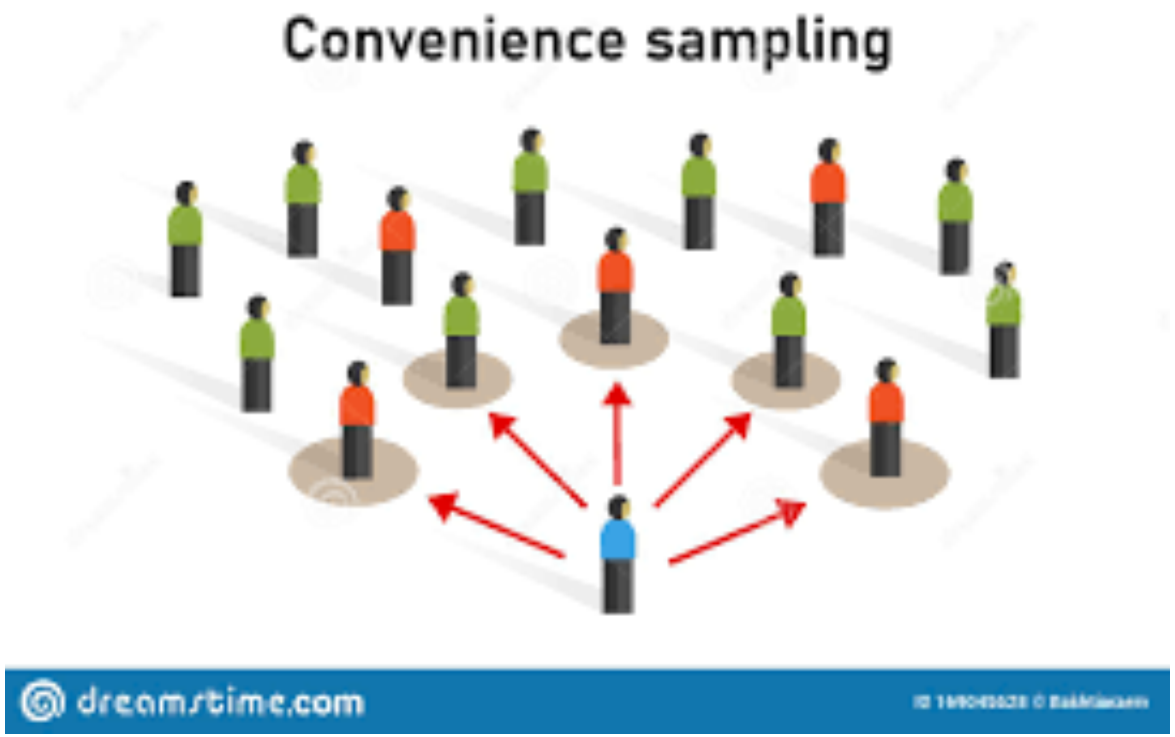
3) Cluster → classrooms within schools are the clusters and we census in each room

Bad Sampling Method

<u>Convenience Sample</u>

- Include individuals who are convenient to sample
- The group may not be representative of the population
- Examples: pollsters in shopping malls, internet polls
- Beware of confounding variables

- This is a sampling technique to AVOID! Bad Sampling that tends to **Bias** results



http://researcharticles.com/index.php/convenience-sampling-in-qualitative-research/

Bias and Variability

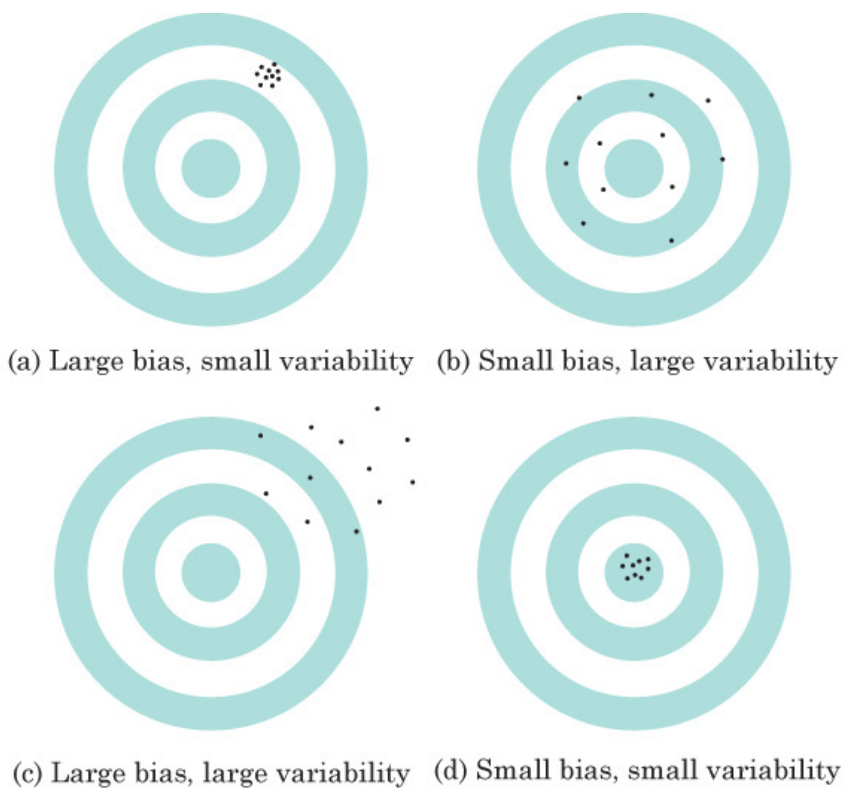
UNBIASED

Bias

- Studies are **biased** if they consistently underestimates or consistently overestimates the true value of the characteristic it measures.

Variability

- Refers to how consistent results are.



(a) Large bias, small variability (b) Small bias, large variability

(c) Large bias, large variability (d) Small bias, small variability

Figure 3.3
Moore/Notz, *Statistics: Concepts and Controversies*, 9e, © 2017 W. H. Freeman and Company

Another Bad Sampling Method

Voluntary Response Sample

- A large group of individuals are invited to respond and all who do respond are counted
- **Those who volunteer generally feel more strongly than those in general population**

- Opinions of a volunteer response sample may be very different from the population as a whole
- Usually only get either STRONG positive or STRONG negative opinions

- Examples: Call-in shows, text-in polls, internet polls, etc…

- This is another sampling technique to AVOID! Bad Sampling that tends to **Bias** results

Errors in Sampling (Sources of Bias)

Random Sampling Error

- The different results from sample to sample.
- Caused by **chance** in selecting a **random sample**.
- **This is natural and always present!**

Sampling Errors

- Errors caused by the **act of taking a sample**.
- They cause sample results to be different from the results of a census.

Nonsampling Errors

- Errors not **related to the act of selecting a sample** from the population. They can be present even in a census.

Misuse of Survey Results

Errors in Sampling (Sources of Bias)

Sampling Errors

Convenience Sampling

- Under or over represents certain groups of a population.

Volunteer Response Sampling

- Only receive opinions that are usually stronger (either negatively or positively) that don't reflect opinions of the population as a whole.

Undercoverage

- Occurs when certain groups of the population are left out.
- An incomplete sampling frame can cause **undercoverage.**
- Example:
    - Only households are contacted, so students in dormitories, prison inmates, most members of the armed forces, the homeless, and people staying in shelters are left out.
    - Many polls interview only in English, which leaves some immigrant households out of their samples.

Overcoverage

- Including people outside of our population of interest in the sampling frame.
- Simple example: Only interested in Freshman and Sophomores, but some Seniors sneak into our list of students to ask.

Errors in Sampling (Sources of Bias)

Non Sampling Errors

Poorly Worded Questions

- ○ Can be based on how a question is worded (guiding people to answer a particular way)

Response Errors

- ○ Incorrect or untrue responses (for whatever reason: lying, bad memory, etc.).
- ○ Ex. Asking someone face to face their GPA in proximity of others.
- ○ Social Acceptability Bias
    - ○ People are reluctant to admit to behavior that may reflect negatively on them.

Nonresponse Error

- ○ Lack of participation, i.e. failure to obtain data from individuals that are selected for the sample
- ○ Missing data

Processing Errors

- ○ Mistakes in mechanical tasks such as arithmetic or data entry.

Misuses of Survey Results

- Generalizing Results - generalizing survey results beyond the scope of the actual data setting (lab rats do not respond the same as people)
- Self Interest Bias - People who have an interest in the outcome of an experiment have an incentive to use biased methods
    - Researchers may dismiss rare negative findings if the company that is paying for the study would be adversely affected

Summary of Sampling Methods

Why we use other sampling methods

- SRS is the simplest method, but…
- Because each sample of size *n* has the same chance of being selected, we cannot obtain information for separate groups of individuals
    - (e.g., individuals of different gender, race, income class, or religion).
    - Our selected sample might miss them….
- So to ensure we get data from each group of interest in our sample…
    - We use methods like Stratified Random Samples and Cluster Samples

GOAL

- Goal of these good methods and good survey techniques (like well worded questions) is to **reduce bias**.
- Ideally, we want the ONLY ERROR to come from **Random Sampling Error.**
- Of course this is not possible in most cases, so we try to **minimize** other sources of errors in our samples.

PROBLEM SESSION!!!!!!!!!!!

Problem #5

As discussed in the chapter, GfK Roper Consulting conducts a global consumer survey to help multinational companies understand different consumer attitudes throughout the world. In India, the researchers interviewed 1000 people ages 13-65. There samples is designed so that they get 500 males and 500 females.

a) Are they using a simple random sample? How do you know?

b) What kind of design do you think they are using?

Problem #5 Solution

a) No, because they have an equal number of males and females.

b) Stratified random sampling, stratified by gender

Problem #9

A business magazine mailed a questionnaire to the human resource directors of all Fortune 500 companies, and received response from 23% of them. Those responding reported that they did not find that such surveys intruded significantly on their workday. Identify the following, if possible. (If not, say why.)

a) The population
b) The population parameter of interest
c) The sampling frame
d) The sample
e) The sampling method, including whether or not randomization was employed.

Problem #9 Solution

a) Human resource directors of Fortune 500 companies
b) The proportion of those who don’t feel surveys intrude significantly on their workday.
c) Human resource directors at Fortune 500 companies
d) The 23% who responded
e) Attempted census – nonrandom
f) Bias – hard to generalize because who responded is related to the question itself (nonresponse bias)

Problem #13

An intern is working for Pacific TV (PTV), a small cable and internet provider, and has proposed some questions that might be used in the survey to assess whether customers are willing to pay $50 for a new service.

Question 1: If PTV offered state-of-the-art high-speed Internet service for $50 per month, would you subscribe to the service?

Question 2: Would you find $50 per month –less than the cost of a daily cappuccino– an appropriate price for high-speed Internet service?

a) Do you think these are appropriately worded questions? Why or why not?
b) Which one has a more neutral wording? Explain.

Problem #13 Solution

a) Question 1 seems appropriate. However, Question 2 predisposes the respondent to answer the question in the affirmative.

b) Question 1 is more neutral in wording. Question 2 is biased; a leading question in which respondents are encouraged or led to answer, “yes.”

Problem #17

For your marketing class, you’d like to take a survey from a sample of all the Catholic Church members in you city to assess the market for a DVD about Pope Francis’s first year as pope. A list of churches shows 17 Catholic churches within the city limits. Rather than try to obtain a list of all members of these churches, you decide to pick 3 churches at random. For those churches, you’ll ask to get a list of all current members and contact 100 members at random.

a) What kind of design have you used?
b) What could go wrong with the design you have proposed?

Problem #17 Solution

a) Multistage design; first you select clusters (the churches) and then you employ simple random sampling to contact 100 members.

b) If any of the 3 selected churches is not representative of the population then you will introduce bias into your sample by including that church in your sample.

Problem #33

A local cable TV company, Pacific TV (PTV), with customers in 15 towns is considering offering high-speed Internet service on its cable lines. Before launching the new service they want to find out whether customers would pay the $75 per month that they plan to charge. An intern has prepared several alternative plans for assessing customer demand. For each, indicate what kind of sampling strategy is involved and what (if any) biases might result.

a) Put a big ad in the newspaper asking people to log their opinions on the PTV website.

b) Randomly select one of the towns and contact every cable subscriber by phone.

c) Send a survey to each customer and ask them to fill it out and return it.

d) Randomly select 20 customers from each town. Send them a survey, and follow up with a phone call if they do not return the survey within a week.

Problem #33 Solution

a) Volunteer sample – only those with strong opinions will respond

b) Cluster sample – if the town is not representative of the population, then the sample will be biased

c) Census – plagued by nonresponse bias – those who respond will differ from those who don’t

d) Stratified random sample – stratified by town, with follow-up. Should be unbiased.