Chapter 11

Sample Surveys

Chapter 11

Homework

Read Chpt 11

P297 1, 2, 3, 5, 11, 13, 17, 18, 21, 23

Chapter 11

Objective

Identify appropriate methods for collecting data.

Background

- I am certain you have learned in other math classes ways to display and summarize data (bar charts, pie charts, mean, median, etc), but those concepts were presented in simplistic methodology and are limited to examining a specific collection of data (descriptive statistics).
  - Often we wish to make decisions about the future or concepts with which we are unfamiliar. In those cases we may need to go beyond the data in front of us (**infer**) and/or consider a much larger world of interest (**population**).
  - There are three major ideas that will allow us to make this **inference**…

1. Sampling

Idea 1: Examine a Part of the Whole

- The first idea is to obtain a **sample**. This is a simple idea but most assuredly, not easily properly accomplished.
- We would like to know about an entire **population** of data values, but examining all of the data is usually impractical, if not impossible.
- We then settle for examining a **subset** of data values (a **sample**) selected from the **population** of interest.
- You apply **sampling** when you take a small taste of sausage at Costco and infer that all the sausages will taste that good.

Idea 1: Examine a Part of the Whole

- Opinion polls are examples of **sample surveys.** Surveys ask questions of a small **representative** group of people in the hope of learning something about the entire population.
  - Professional pollsters work very hard to ensure that the **sample** taken is **representative** of the **population**.
  - If a sample is not truly **representative** of the entire **population**, that **sample** may suggest erroneous information about the population.
  - In the early days of computing that was known as "garbage in, garbage out".

GIGO

BIAS

- Sampling that is **not representative** of the population (tends to over- or under-emphasize some characteristic or quality of the **population**) is said to be **biased**.
  - **Biased** simply means "the sample is **not representative** of the **population**". **Bias** is the Simon Legree of statistics—the one thing statisticians work very hard to avoid.
  - There is no way to fix a **biased** sample and no way to salvage useful information from that biased sample (GIGO).
- The best way to avoid **bias** is to select individuals for inclusion in the sample …

RANDOMLY!!

2. Randomization

Idea 2: Randomize the selection of a sample

- **Randomization** can inoculate researchers against factors that you know are in the data and may influence the data.
- **Randomization** can also help protect against factors of which you are blissfully unaware (lurking or confounding variables).
  - Randomizing selection ensures that, to the extent possible, the sample looks like the rest of the **population** (**representative**). A pequena poblacion.

Idea 2: Randomize the selection of a sample

- In addition to protecting against bias, random selection also makes it possible for us to **infer** about the **population** when we have access to only a sample. **Inferential statistics**.
  - **Caution**, **inferring** from a sample is only made possible when we are careful to choose data values **randomly** with appropriate techniques.

- I have used the term **population** several times. Keep in mind that, for us, a **population** is a collection of observations (data). We are interested in characteristics of individuals, not the individuals themselves. Yes, sometimes you are just a number.

3. Size Matters

Idea 3: It is the size of the sample!

- How large a **random sample** do we need for the **sample** to be reasonably representative of the **population**?
  - **It is the size of the sample**, not the size of the population, that makes the difference in **sampling**.
    - **Exception**: If the **population** is small enough and the **sample** is more than 10% of the whole **population**, the population size can matter.
  - The fraction of the **population** that you have sampled does not matter. It is the **sample size** that is important.

Getting the chili juuuust right

- Suppose you are making chili for a chili cook-off.
  - If you use too small a spoon to check the flavor, you will not get a good taste of the chili.
  - If you use too large a spoon to check the flavor, you will not have enough left to enter the judging.
  - So you need a spoon that is juuuuust right.
  - It does not matter how large the chili pot, just how large the spoon.

What if We Could Get to Everyone?

- Why bother determining the right sample size? Why not simply include everyone in the population?
  - Such a special sample is called a **census**.
  - How do you get to everyone? It can be difficult, if not impossible to complete a census. The United States tries every 10 years. Look up how that works out. There will nearly always be some data values that are hard (or expensive) to locate or difficult to measure
  - It would take so long, the population would change during the attempt.

Population Parameters

- **Models** use mathematics to represent reality. Models are the ideal distribution. (Math is everything, everything is math.)
  - **Parameters** are key values in **models**.
  - A **parameter** that is part of a **model** is called a **population parameter**. An example of a **population parameter** is the **population** mean.
    - We use sample data to **estimate population parameter**.
- **A summary** found from sampling data is a **statistic**.
  - The values derived from sample data that estimate **population parameters** are called **sample statistics**.

Notation

- We typically use Greek letters to represent **population parameters** and Latin (English) letters to denote **sample statistics**.

Name	Statistic	Parameter
Mean	$\bar{Y}$	μ mu
Standard Deviation	S	σ sigma
Correlation	r	ρ rho
Regression	b	β beta
Proportion	$\hat{P}$	P

Simple Random Sample

- Your book uses **SRS** for **S**imple **R**andom **S**ample. I prefer you write it out.

- Our goal is to ensure the **statistics** we compute from the **sample** accurately reflect the corresponding **population parameters**.

  - Once again: *A* **sample** that accurately represents the **population** from which that sample was drawn is said to be **representative** of the population.

Simple Random Sample

- Random selection requires that each **individual observation** in the **population** has an **equal chance of being included in the sample**.
  - It is also necessary that every possible **sample** of the size we plan to obtain has an **equal chance of being the sample selected**.
  - In addition to each value, with truly random sampling, each **combination** of values (**sample**) has an equal chance of being selected.
- This method is called a **Simple Random Sample (SRS)**.
  - A **Simple Random Sample** is the gold standard against which we measure other sampling methods, and the method on which the theory of working with sampled data is based. Sampling is not perfect, but it works!

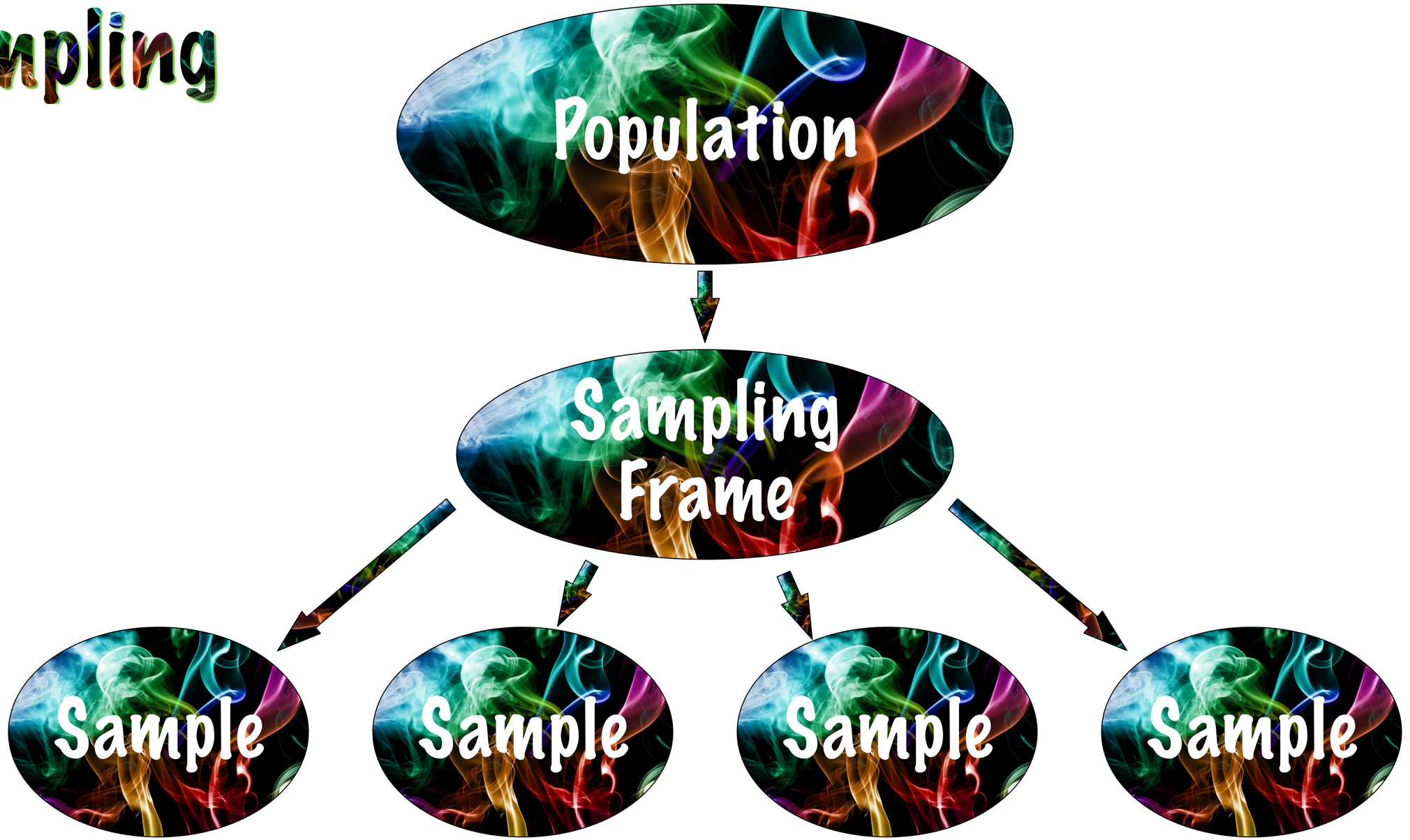
Simple Random Sample

- To select a sample at random, we first need to define from whence the sample will come.
  - The **sampling frame** is a collection of observations from which the sample can actually be drawn. The **sampling frame** is a subset of the **population**.
  - From the **sampling frame**, the theoretical methodology for choosing a **Simple Random Sample** is to record every data value in the **sampling frame** on identical papers, put all the papers in a hat, mix, and blind draw the quantity desired.
  - An alternative is to assign a random identification number to each data value in the **sampling frame** and use a random number generator to select the desired number of data values for the sample.

Simple Random Sample

- Here is where statistics gets interesting. Different samples drawn at random from the **same sampling frame** are expected to be different.
  - Each draw of random numbers results in **different** data for that sample.
  - These differences lead to **different** values for the variables we measure (e.g. mean).
  - We call these sample-to-sample differences **sampling variability (or sampling error).**
  - It is this **sampling variability** that forms the basis for statistics. **Sampling variability** is an extremely important concept. Get it straight in your head.

Sampling



Other Random Samples

- **Simple Random Sampling** does have some issues. Actually the attempt to truly select randomly brings difficulties. Fortunately simple random sampling is not the only fair way to sample.
  - Designs used to sample from large populations are often more complicated than the **Simple Random Samples**.
  - All statistical sampling designs have in common the idea that **chance**, rather than human choice, is used to select the **sample**. Thus, **there is an element of randomness in every method**.
  - These other techniques of random sampling are designed to overcome problems of which the statistician is aware (or maybe unaware).

Stratified Random Sampling

- Sometimes the population is first divided into **homogeneous** groups, called **strata**, before the sample is selected.

  1. Divide the population by some important characteristic to attempt to balance the sample. Doing so ensures our sample contains the desired distribution of that characteristic. (e.g. male/female, white/latinx/black/asian, etc…)

  2. Then we simply use simple random sampling **within** each **stratum** and combine the results to create our sample.

- This common sampling design is called **stratified random sampling**.

Stratified Random Sampling

- For instance; suppose we wish to choose a sample from our population of interest that is the students from CHHS.
  - If we believe that the class of the student (9,10,11,12) might have an influence on the results, we will want to ensure our sample has an appropriate number of Seniors, Juniors, Sophomores, and Freshmen.
  - To ensure appropriate representation, we randomly select the correct number from each class (**strata**) to accurately reflect the ratios in the population.
  - The end result is a sample with the same proportion of each class as we expect to find in the population.

Stratified Random Sampling

- The most important benefit of stratifying is that stratifying can **reduce the variability** of our results.
  - When we restrict by strata, other samples are more like one another, so statistics calculated will vary less from one sample to another.
  - Stratified random sampling can reduce bias by accounting for the variability created by the characteristic determining the strata.

Cluster Sampling

- Sometimes stratifying isn't practical and simple random sampling is difficult or impossible.
- Dividing the population into similar parts or **clusters** can make sampling more practical. Usually the clusters are found in the population and are not created by the researcher.
  - Then we could select one or a few clusters **at random** and perform a **census** within each of them.
    - This sampling design is called **cluster sampling**.
  - **If each cluster fairly represents the full population**, cluster sampling will give us an unbiased sample.

Cluster Sampling

- In our CHHS example; instead of class strata, we might cluster by selecting quads, if we believe each quad is a good representation of the entire population of CHHS, then using cluster sampling would be reasonable.
  - Within each quad we could obtain data from the students (census).

Cluster vs. Stratified Sampling

Cluster sampling is not the same as stratified sampling.

- We stratify to ensure that our sample represents different groups in the population, and sample randomly within each stratum.
- Strata are **internally homogeneous** on some characteristic, but the strata are different from each other on that characteristic.
- Clusters are more or less alike, are **internally heterogeneous** and **each resembles the overall population**.
  - We select clusters to make sampling more practical or affordable.
  - Usually clusters are naturally occurring.

Cluster vs. Stratified Sampling

- Your book has a good metaphor for the difference between cluster sampling and stratified sampling.
  - A layer cake has layers of cake separated by icing or custard.
    - **Stratified sampling** would separate the layers and **randomly** select samples from each layer to test.
    - **Cluster sampling** would cut the cake in slices, **randomly** select a few slices, and eat each slice in its entirety.

Cluster vs Stratified Sampling

- In **Stratified sampling** we would separate the cake into the individual layers and randomly choose an appropriate amount from each layer for our sample.

- In **Cluster sampling** we randomly select a slice or several slices, and eat the whole slice as our sample.

Systematic Samples

- Sometimes we draw a sample by selecting data values **systematically**.
  - For example, you might survey every 10th person on an alphabetical list of students.
- To make it **random**, you must still start the systematic selection from a **randomly** selected data value.
- If there is no reason to believe that the order of the list is associated with the variables studied, **systematic sampling** can result in a representative sample.
- The primary advantages of **systematic sampling** are cost and ease of selection.

Multistage Sampling

- Sometimes we use a variety of sampling methods together.
  - For instance, we might choose (1) cut our cake into slices, then (2) stratify by layers.
- Sampling schemes that combine several methods are called **multistage sampling**.
  - For our school, we could cluster by quad, cluster by classrooms in the quad, then randomly sample from the students in those classrooms.
- Most surveys conducted by professional polling organizations use some combination of stratified and cluster sampling as well as simple random sampling.

Defining the "Who"

- When collecting data you need to start by determining the population of interest. That may be more difficult than you expect. Even if you can accurately define the population of interest (Who) it may be difficult to collect data from them.

- That difficulty leads to the **Sampling Frame**. This is not a complete collection of our population but it is the best we can do. Keep that in mind when interpreting results.

- Finally we get to our target sample. It is from the sample we get our data. Not every member of the sample (remember, it was randomly generated) is going to be eager to participate. Non-response can be an issue in sampling.

Defining the "Who"

- At each step, the group we can study may be constrained somewhat.
  - With each constraint the *Who* may change and can introduce bias.
- When analyzing data a careful study will address the question of how well each group matches the population of interest. Sounds simple, but ain't easily accomplished.
  - In other words, each time we limit the choices, the greater the chance that our sample loses accuracy in representing the population.

Defining the "Who"

- One of the main benefits of a true simple random sample is that it avoids the constraints and never loses our *Who*.
  - Since simple random sampling involves every member of the population having equal opportunity to be included in the sample, and every combination of members of the population is equally likely, we are more confident in the sample being representative of the population.

The Valid Survey

- A **valid** survey provides desired information about a population. It is not sufficient to just grab a sample and then start asking questions. Before doing anything, a researcher must ask herself:
  - What do I want to know?
  - Am I asking the right respondents?
  - Am I asking the right questions?
  - What would I do with the answers if I had them?
  - And would they truly address the things I want to know?
- It sounds simple but you will be amazed by how difficult that can be.

The Valid Survey

- The questions may sound obvious, but there are a number of pitfalls to avoid.
  - Know what you want to know.
    - Know what you hope to learn and from whom you hope to learn it. When you ask someone to go to the football game with you, what is it you really want to know? Going to the game, or going with you?
  - Use the right frame.
    - Be sure you have a suitable **sampling frame**. Asking patients which surgery is most appropriate is probably not the best sampling frame.

The Valid Survey

- Make your questions specific rather than general.

- Design a valid instrument.
  - The instrument itself can be the source of error (Response Bias). Too long a survey yields fewer responses, ambiguous questions yield meaningless responses.

- Do you eat breakfast in the morning?
  - What constitutes breakfast? A piece of toast?
  - Each morning? Every morning? Most mornings?

The Valid Survey

- Ask for quantitative results when possible.
  - On a scale of 1 to 5 how important is eating something in the morning?
- Be careful in phrasing questions.
  - A respondent may not understand the question or may understand the question differently than the way the researcher intended.
  - Even subtle differences in phrasing can make a difference.
    - Did you eat this morning?
    - Did you eat breakfast this morning?
- It is usually a better idea to offer choices rather than inviting a free response.

How to Sample Badly

- **Voluntary Response Bias**
  - In a **voluntary response sample**, a large group of individuals is invited to respond, and only those who do respond are counted.
    - Voluntary response samples are **almost always biased**, and conclusions drawn from them are almost always useless.
    - Online Facebook survey.
  - Voluntary response samples are almost certainly biased toward those with strong opinions or those who are strongly motivated. In other words, not representative of the population.
  - Since the sample is not representative, the resulting **voluntary response bias** invalidates the survey.

How to Sample Badly

- **Convenience Sample**
  - In **convenience sampling**, we simply include individuals who are convenient.
    - Unfortunately, this group is unlikely to be representative of the population.
- Convenience sampling is not only a problem for students of statistics.
  - It is a widespread problem in the business world and in media. The easiest people for a company or media outlet to sample are its own clients. Those subjects are not likely to be representative of the population.

How to Sample Badly

- **Sample from a Bad Sampling Frame:**
  - A simple random sample from an incomplete sampling frame creates bias because the individuals not included in the sampling frame (under-coverage) may differ from the ones that are not included in the sampling frame and thus not in the sample..

- **Under-coverage (Underrepresented):**
  - Many bad survey designs suffer from **under-coverage**, in which some sector of the population whose representation (underrepresented) in the sampling frame, and/or the sample, is less than what that sector actually represents in the full population.

What Else Can Go Wrong?

- **Watch out for those in the sample choosing not to participate (nonrespondents).**
  - A common and serious potential source of bias for most surveys is **nonresponse bias.** **Remember the sample is chosen randomly before actual data collection.**
  - No survey succeeds in getting responses from everyone.
    - The problem is that those who do not respond may differ from those who do, and the difference may affect the variables and results we are seeking.

What Else Can Go Wrong?

- Don't ask respondents to participate in surveys that go on for too long.
  - Surveys that are too long are more likely to be refused or incomplete, reducing the response rate and biasing **all** the results. The attention span of most respondents is not very long.

- Be extremely careful to avoid influencing the responses yourself.
  - **Response bias** refers to anything **in the survey** design that might influence the responses.
    - For example, the wording of a question can influence the responses:
      - "The cars collided" or "the cars smashed into each other".

Response Bias

- **Response bias** refers to anything **within the survey** or the way the survey is conducted that might influence the responses. Response bias has nothing to do with the respondents.

Terms to Remember

- Voluntary Response Bias
- Convenience Sample
- Bad Sampling Frame
- Under-coverage
- Nonresponse Bias
- Response Bias

And Why.

- Bias can also arise from poor sampling methods:
  - **Voluntary response bias** stems from samples that are allowed to select themselves.
  - **Convenience samples** are likely to be flawed as they are unlikely to be representative of the population
  - **Under-coverage** occurs when individuals from a subgroup of the population are selected less often than they should be (under-represented).

And Why.

- **Bad Sampling Frame** arises when the parts of the population that is actually accessible is not a good representation of the population.

- **Non-response bias** can arise when sampled individuals will not or cannot respond.

- **Response bias** arises when respondents' answers might be affected by external influences, such as question wording or interviewer behavior.