Week 5: Data Collection

Jessie Yeung

STA 220

Fall 2024

Agenda for today

1. Cover the topic of data collection
2. R content
3. Discuss Term Test [redacted] next week

Overview

- This week will cover Module 5
  - Module 5: Data Collection
- Topics include:
  - Different types of Sampling Procedures
  - Observational Studies
  - Experimental Studies

Types of Sampling Procedures

Why must we collect data?

- When we conduct research in any field, we are attempting to answer some research question.
- In general, we can't conduct our research on everyone/everything because it's often to expensive, time-consuming, impractical, etc.
- So we want to take the results/conclusions that we make from our research and apply it in a more general setting (i.e. to everyone/everything)
  - this is called **statistical inference** Result from sample ⟶ Apply to larger Population
  - we want to determine whether the result we see in our research is due to chance or is actually the true result in the real world.
- This will require some form of data collection to investigate our research hypothesis.

Inference and Sampling

- Since the goal of the study is to look at a small amount of data and hope that we can generalize it, we need to make sure we know who/what we want to generalize to
- Therefore, before even collecting data, we need to define our **population of interest.**
    - these will be all the people/things that we are trying to learn something about
    - this is not the same thing as all people/things everywhere
    - for example, what if I want to see if a new drug will make a headache go away faster than the old drug.
        - my population of people won't be all people everywhere
        - why would I give the drug to someone that doesn't have a headache…
        - in this case, my population would be all people who have a headache

Recall: Gluten-free diet example

- Back in the first week of class, we discussed a study that tried to determine whether a gluten-free diet (GFD) for people without gluten sensitivities was beneficial to their gastrointestinal health.
- Ran a double-blind randomized trial over a 2 week period
  - half of the 28 subjects used high gluten flour in the 2 week period, while otherwise keeping to their GFD
  - other half used gluten-free flour alongside their normal GFD
- Various indicators of GI health measured before the trial and again at the end
- Result: GI health was statistically the same, regardless of whether there was gluten in the diet or not!

Recall: Gluten-free diet example

In the GFD study, what is my population of interest?

a) All people on gluten free diets

b) All people in the world

[✓ c)] All people without gluten sensitivities

d) All people with GI health issues

e) All people in the study

Data from the Population

- For most types of studies, it is not feasible to collect information/data from every member of the population of interest.
- If we did this, we would be conducting a **census**.
    - in the GFD study, we would need to collect data from everyone who doesn’t have gluten sensitivities in the entire world!
    - we know that 1% of Canadians have celiac disease, and 6% have gluten sensitivities as diagnosed by a physician
    - so we would need to collect information from 93% of Canadians
        - that’s over 37 million people!
        - and that’s just Canadians!

Alternative to a Census

- When a census is not possible (which is most of the time), then we have to collect information on a smaller portion of the population of interest.
    - this is called the **sample**
    - we still want to talk about the population, but now using the results from the sample
- If we take a census, then we don't have to worry about randomness because we have collected everything.
    - when we take a sample, now randomness has to be considered.
    - depending on who is in our sample, we will calculate different sample statistics (like sample mean, sample proportion, etc.)
    - need to consider how likely we are to have chosen that sample and therefore gotten that sample statistic.

What are we measuring?

- Whether we are dealing with a sample or the entire population, we are always trying to measure something.
- When we collect a sample, we calculate a **sample statistic** (e.g. sample mean)
- We are attempting to use this statistic to make a guess (or make inference) about the corresponding value in the population
  - this value in the population is called a **population parameter** (e.g. population average)
- When we use our sample statistic to make inference about a population parameter, we call this estimation
  - we calculate a statistic to estimate an unknown parameter
  - this is because we did not take a census, so we do not know this population parameter.

Parameter vs. Statistic

- Parameters (or population parameters) are unknown values about the population.
  - They are unknown to us unless we take a census with perfect measurement
  - These are the values that are of interest to us!
- Statistics (or sample statistics) are summary measures we can calculate based on our sample
  - We can calculate them and obtain a number
  - They differ from sample to sample
- Goal: Use sample statistics to estimate the unknown population parameters
  - Estimate the population mean μ using sample mean $\bar{x}$
  - Estimate the population proportion p using the sample proportion $\hat{p}$

Check your understanding!

In the GFD study, we are told "*various indicators of GI health were measured and summarized*". Are these statistics or parameters?

[✓ a)] Sample statistics

b) Population parameters

What makes a good sample?

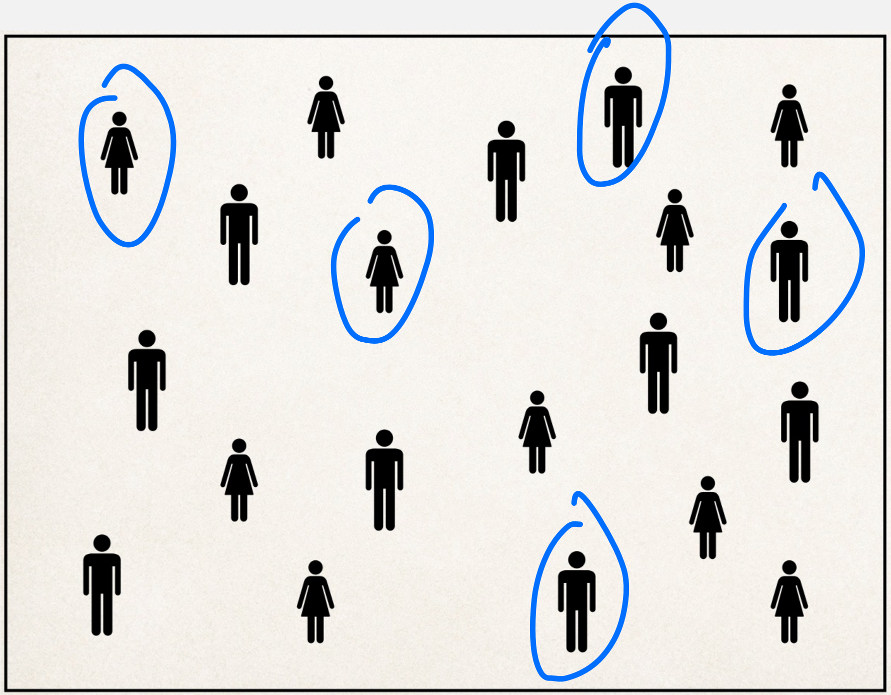
- The properties that make a sample good and therefore make the sample statistic a good estimate for the population parameter are:
  - the sample should be representative
    - this means that the people/things in the sample should have the same characteristics as the people/things in the population
    - the sample should basically look like a mini population
  - the sample should also be taken randomly
    - this says that all people/things in the population have the same chances of ending up in the sample
    - helps make sure that the sample is representative

How to get a sample?

- We have a number of different methods that we can use to select a sample from some population.
- Each of these ensure that our sample will be selected randomly and will be representative of the population.
    - **Simple Random Sampling (SRS)**
        - each person in population has equal chance of being sampled
    - **Stratified Sampling**
        - divide population into non-overlapping subgroups (strata), then take SRS in each subgroup
    - **Cluster Sampling**
        - divide population into non-overlapping subgroups (clusters), then take SRS of the clusters, including everyone in the sampled clusters

Simple Random Sampling

Simple Random Sampling (SRS): each person in population has equal chance of being sampled

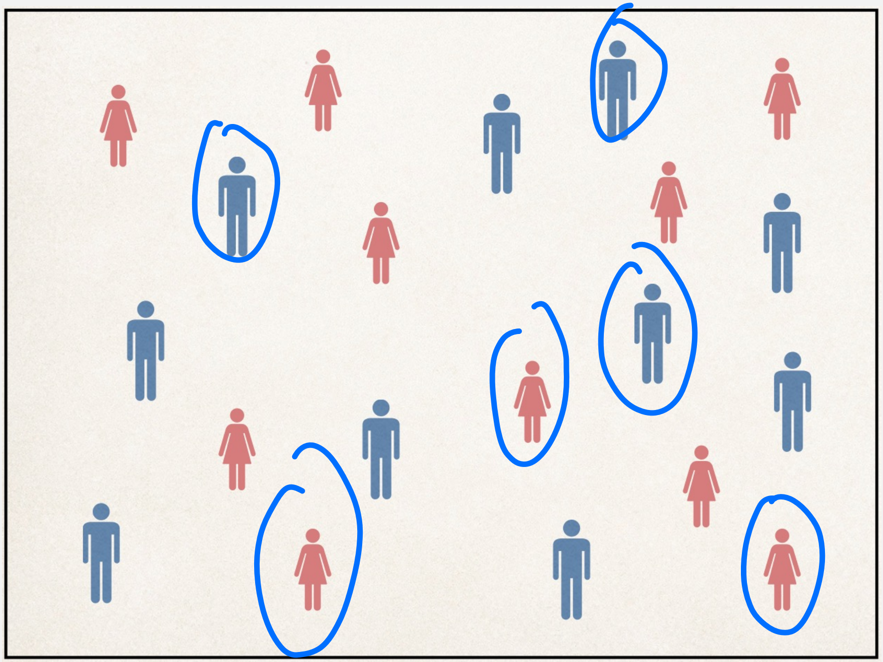


Population

We have 20 people in the population. Let's take a sample of 5 people.

Stratified Sampling

Stratified Sampling: divide population into non-overlapping subgroups (strata), then take SRS in each subgroup

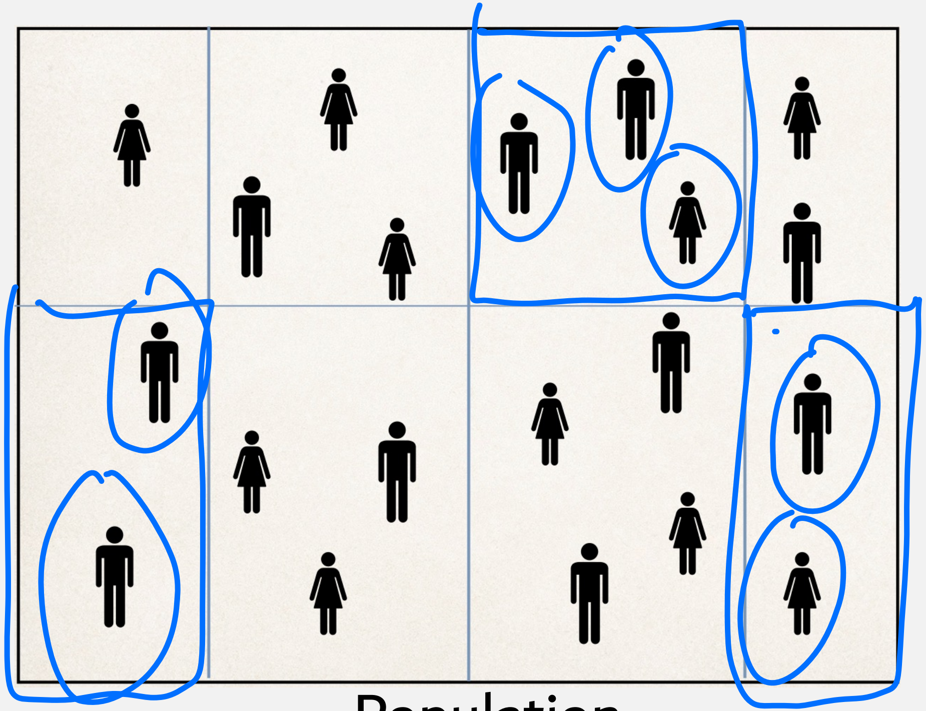


Population

We have 2 strata (men and women). Let's take a sample of 3 people in each group:

Cluster Sampling

Cluster Sampling: divide population into non-overlapping subgroups (clusters), then take SRS of the clusters, including everyone in the sampled clusters



Population

We have 8 clusters in the population. Let's take a sample of 3 of them:

Which method should I choose?

- Simple Random Sampling is the simplest option.
- When you have some sort of natural grouping, you should use Stratified:
  - groupings like gender, by province or state, etc.
  - ends up being more convenient and means you will always sample someone with each of these characteristics.
- When it is easier to select groups rather than individuals, you should use Cluster:
  - usually easier implies less expensive, means you won’t need to travel as far or hire as much help to get a sample.
  - again, groupings end up being fairly natural, such as schools, classes, postal codes, etc.
  - only want to use this if you want the whole group sampled.

Other Sampling Methods

- The previous 3 sampling methods will provide samples that are random and representative of the population being sampled.
- There are two other methods that are not random methods of sampling:
- **Systematic Sampling**
  - take a list of the population and select every k^{th} person, where the starting position is chosen randomly
- **Convenience/Volunteer Sampling**
  - use the first n individuals who agree to participate.
- Unrepresentative samples can sometimes occur with Systematic sampling, but will always occur with Convenience Sampling

Check your understanding!

In the GFD study, what type of sampling was used?

> *“Participants (who received no financial incentives), recruited by advertising, were ≥18 years, had no diagnosed gluten-related disorders, and followed gluten-containing diets.”*

a) Simple Random Sampling

b) Stratified Sampling

c) Cluster Sampling

d) Systematic Sampling

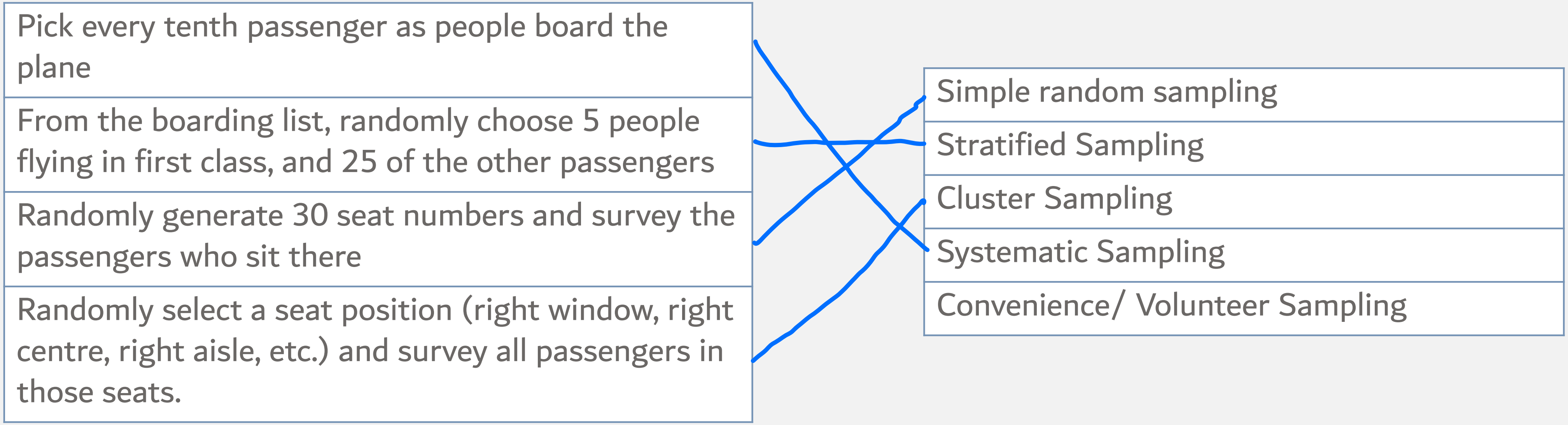
[✓ e) Convenience/Volunteer Sampling]

Unrepresentative Samples - Problems

- When our sample is not representative of the population, we introduce **bias** into our results.
- This means that when we try to use our sample statistics to estimate the population parameter, we don't get a very accurate guess.
  - the mean of the sampling distribution is not the same as the mean of the individual observation distribution. $E(\overline{X}) = \mu$
- Types of bias that we can see:
  - Selection bias: happens when we systematically exclude portions of the population from our sample.
  - Measurement/Response bias: the way data are collected creates a systematic difference between the observed values and the true values.
    - The way a question is written or how an interviewer acts or other conditions impact the response.
  - Nonresponse bias: when we don't collect data from people in our sample

Example: Passengers on a Flight

We need to survey a random sample of 300 passengers on a flight from Vancouver to Tokyo. Match he following statements to the sampling method being described.



Pick every tenth passenger as people board the plane

From the boarding list, randomly choose 5 people flying in first class, and 25 of the other passengers

Randomly generate 30 seat numbers and survey the passengers who sit there

Randomly select a seat position (right window, right centre, right aisle, etc.) and survey all passengers in those seats.

Simple random sampling

Stratified Sampling

Cluster Sampling

Systematic Sampling

Convenience/ Volunteer Sampling

Observational and Experimental Studies

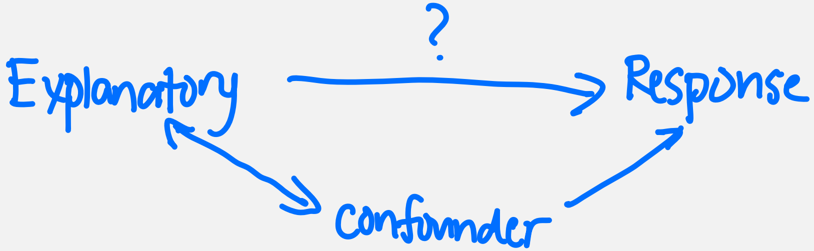
Types of Studies

- When we collect data from a sample, we often look at multiple variables at a time.
- We therefore often want to determine whether there is an association between certain variables that is not due to chance.
- In particular, we may be interested to determine if we can say that one variable causes another variable.
- We therefore need to choose the right study type, in addition to a good sample, in order to make these conclusions

Types of Studies

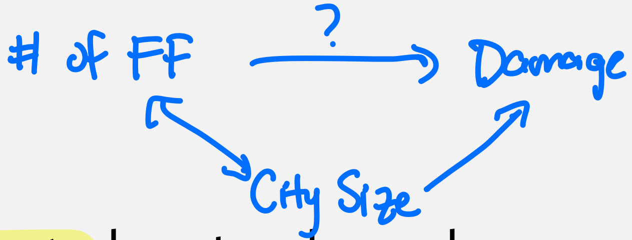
- There are two main types of studies, each which has different useful features.
- **Observational studies**: useful for drawing conclusions about the population, or about differences between two or more groups, or relationships between variables
  - No control over who is in which group - you are just observing
- **Experimental studies**: useful for drawing conclusions about the cause of a certain variable
  - Able to control who belongs in which group - you get to intervene

Terminology in Studies



- The variable that is the main interest of the study is called the response variable or the outcome variable.
  - it is the variable that responds to what is happening with other variables
- The variables that we are trying to use to explain what is happening in the response variable are called the explanatory variables.
  - however, it is important to distinguish between the variables that can realistically explain the response, and those that confound the relationship between other variables and the response, called confounders
  - A confounding variable is one that is associated with the explanatory variable.
  - Means we have to be very careful about how we collect data and what variables we use

Example: Confounding Variable



- Suppose I'm still looking at the association between damage done to a house by a fire and the number of firefighters, but I collected my data across all of Ontario, which means that some house fires are in small towns and others in big cities.
- Being in a small town is associated with fewer firefighters
    - Because of this, I may get more firefighters at house fires in cities but fewer firefighters in small towns.
    - So whether or not we are in a small town or big city affects my explanatory variable
- But being in a small town may also mean that houses are made of more combustible material, so fires will spread faster
    - Thus living in a small town may also have an effect on the response.
    - so location of fire is a confounding variable.
    - we don't know if the amount of damage we see is explained by the number of firefighters or by whether the fire was in a small town.

Observational vs Experimental

- So which type of study should I choose?
- It all depends on your resources and the type of question you want to study.
- Observational studies involve collecting information on both the explanatory and response variable, and seeing if there's a relationship
  - Used if the explanatory variable cannot be manipulated or changed by the study
  - e.g. identify people with a disease (explanatory) and follow them up to see how long they live, or look back at their medical history (response)
    - either way, you cannot control anything about the people or their life
    - means you can never say that something caused anything
  - Cannot conclude that the explanatory variable caused the response outcome

Observational vs Experimental

- Experimental studies involve changing participant’s explanatory variables and then measuring the response variable
  - Use when main interest is to determine whether an explanatory variable *causes* the response
  - This is because it is possible to control what values of the explanatory variable people get, so by giving them one value you can directly see if it causes the response.
  - Means we can randomize our subjects to each value of the explanatory variable
  - e.g. identify if taking a new drug (explanatory) improves sleep quality (response)
- Obviously, it is not always possible to use an experiment to answer your question.
  - e.g. does smoking cause lung cancer? - I can’t make half of you start smoking.
- Experimental studies also give us the chance to get rid of confounding variables so that we can actually say that one variable causes another without the confusion of the confounder.

Videos and Practice Problems

- In this lecture, we covered all of Module 5: Data Collection
  - Note: The modules describe more terminology than we cover in the lecture. Focus on the terminology that we discussed in lecture
- No practice problems for this week's content

Coming up

- Usual quiz is due on Sunday
- Term test is next week on Feb. 13, on campus, during class time
    - See announcement for further information
    - No lecture after the test
- Reading week is the following week