# Data Collection

Nate Wells

Math 141, 2/15/21

## Outline

In this lecture, we will. . .

# Outline

In this lecture, we will. . .

- Discuss principals of data collection
- Compare and contrast observational studies and random experiments

Section 1

## Principles of Data Collection

## Populations and Samples

- Every statistical investigation must begin by clearly identifying the **population** to be studied, the **variables** to be measured, and the **sample** from which measurements will be taken.

# Sampling

*"This is a program that intrudes on people's lives, just like the Environmental Protection Agency or the bank regulators. We're spending $70 per person to fill this out. That's just not cost effective, especially since in the end this is not a scientific survey.* ***It's a random survey.****"*

— D. Webster, Congressman, on the *American Community Survey*

## Sampling

> *"This is a program that intrudes on people's lives, just like the Environmental Protection Agency or the bank regulators. We're spending $70 per person to fill this out. That's just not cost effective, especially since in the end this is not a scientific survey.* ***It's a random survey.****"*
>
> — D. Webster, Congressman, on the *American Community Survey*

- How can a random sample allow us to make justified, scientific conclusions about a population?

# Sampling

> *"This is a program that intrudes on people's lives, just like the Environmental Protection Agency or the bank regulators. We're spending $70 per person to fill this out. That's just not cost effective, especially since in the end this is not a scientific survey.* ***It's a random survey.****"*
>
> — D. Webster, Congressman, on the *American Community Survey*

- How can a random sample allow us to make justified, scientific conclusions about a population?
    - Properties of probability allow us to quantify uncertainty.

# Sampling

> *"This is a program that intrudes on people's lives, just like the Environmental Protection Agency or the bank regulators. We're spending $70 per person to fill this out. That's just not cost effective, especially since in the end this is not a scientific survey.* ***It's a random survey.****"*

— D. Webster, Congressman, on the *American Community Survey*

- How can a random sample allow us to make justified, scientific conclusions about a population?
  - Properties of probability allow us to quantify uncertainty.
  - In isolation, a single random event may seem arbitrary. But in aggregate, a collection of random events is predictable.

## Sampling

> *"This is a program that intrudes on people's lives, just like the Environmental Protection Agency or the bank regulators. We're spending $70 per person to fill this out. That's just not cost effective, especially since in the end this is not a scientific survey.* ***It's a random survey.****"*
>
> — D. Webster, Congressman, on the *American Community Survey*

- How can a random sample allow us to make justified, scientific conclusions about a population?
  - Properties of probability allow us to quantify uncertainty.
  - In isolation, a single random event may seem arbitrary. But in aggregate, a collection of random events is predictable.
- By following basic procedures for randomly selecting a sample, we can be certain that the results fall within a specified margin of the true value a particular percentage of the time.

# Simple Random Sampling

- The most basic form of random sampling is **simple random sampling** (SRS), where every member of the population has an equal chance of being selected for the sample.

## Simple Random Sampling

- The most basic form of random sampling is **simple random sampling** (SRS), where every member of the population has an equal chance of being selected for the sample.
    - Imagine that a unique ID for each individual is written on a slip of paper and thoroughly shuffled togther. IDs are then drawn one-by-one to create a sample.

# Simple Random Sampling

- The most basic form of random sampling is **simple random sampling** (SRS), where every member of the population has an equal chance of being selected for the sample.
  - Imagine that a unique ID for each individual is written on a slip of paper and thoroughly shuffled togther. IDs are then drawn one-by-one to create a sample.
- Importantly, by construction, there is no inherent correlation between any two members of the sample.

# Simple Random Sampling

- The most basic form of random sampling is **simple random sampling** (SRS), where every member of the population has an equal chance of being selected for the sample.
  - Imagine that a unique ID for each individual is written on a slip of paper and thoroughly shuffled togther. IDs are then drawn one-by-one to create a sample.
- Importantly, by construction, there is no inherent correlation between any two members of the sample.
- Its possible a particular sample may not be "representative" of the population (provided it was not caused by systematic error in sampling).

## Simple Random Sampling

- The most basic form of random sampling is **simple random sampling** (SRS), where every member of the population has an equal chance of being selected for the sample.
  - Imagine that a unique ID for each individual is written on a slip of paper and thoroughly shuffled togther. IDs are then drawn one-by-one to create a sample.
- Importantly, by construction, there is no inherent correlation between any two members of the sample.
- Its possible a particular sample may not be "representative" of the population (provided it was not caused by systematic error in sampling).
  - In fact, it is necessary that such underrepresentation samples are possible, in order to quantify *extreme* events.

## Statistial Bias

- Non-random sampling may create **statistical bias**, where certain outcomes are systematically preferred due to sampling technique.

# Statistial Bias

- Non-random sampling may create **statistical bias**, where certain outcomes are systematically preferred due to sampling technique.
- Some particular types of bias include:
    - **Non-response**, where an individual selected for a sample cannot or will not contribute.

# Statistial Bias

- Non-random sampling may create **statistical bias**, where certain outcomes are systematically preferred due to sampling technique.
- Some particular types of bias include:
  - **Non-response**, where an individual selected for a sample cannot or will not contribute.
  - **Undercoverage**, where some groups of a population are less likely to be included in the sample.

# Statistial Bias

- Non-random sampling may create **statistical bias**, where certain outcomes are systematically preferred due to sampling technique.
- Some particular types of bias include:
  - **Non-response**, where an individual selected for a sample cannot or will not contribute.
  - **Undercoverage**, where some groups of a population are less likely to be included in the sample.
  - **Response**, where a sampled individual does not provide accurate or truthful data.

# Statistial Bias

- Non-random sampling may create **statistical bias**, where certain outcomes are systematically preferred due to sampling technique.
- Some particular types of bias include:
    - **Non-response**, where an individual selected for a sample cannot or will not contribute.
    - **Undercoverage**, where some groups of a population are less likely to be included in the sample.
    - **Response**, where a sampled individual does not provide accurate or truthful data.
    - **Self-selection**, where membership in the sample is voluntary (leading to correlation between results and traits promoting participation)

# Statistial Bias

- Non-random sampling may create **statistical bias**, where certain outcomes are systematically preferred due to sampling technique.
- Some particular types of bias include:
  - **Non-response**, where an individual selected for a sample cannot or will not contribute.
  - **Undercoverage**, where some groups of a population are less likely to be included in the sample.
  - **Response**, where a sampled individual does not provide accurate or truthful data.
  - **Self-selection**, where membership in the sample is voluntary (leading to correlation between results and traits promoting participation)
  - **Convenience**, where "randomization" is performed by selecting a convenient block of individuals in the population (leading to strong correlation between members of the sample)

# Sampling Example

- Suppose we want to know how Portlanders feel about a new Thai restaurant in Woodstock.

# Sampling Example

- Suppose we want to know how Portlanders feel about a new Thai restaurant in Woodstock.
- The particular restaurant has a yelp rating of 3.5 stars with 100 reviews.

## Sampling Example

- Suppose we want to know how Portlanders feel about a new Thai restaurant in Woodstock.
- The particular restaurant has a yelp rating of 3.5 stars with 100 reviews.
  - Can we conclude that a typical Portlander would rate this restaurant at 3.5 stars?

## Sampling Example

- Suppose we want to know how Portlanders feel about a new Thai restaurant in Woodstock.
- The particular restaurant has a yelp rating of 3.5 stars with 100 reviews.
  - Can we conclude that a typical Portlander would rate this restaurant at 3.5 stars?
  - What sources of bias are present in this sample?

## Sampling Example

- Suppose we want to know how Portlanders feel about a new Thai restaurant in Woodstock.
- The particular restaurant has a yelp rating of 3.5 stars with 100 reviews.
    - Can we conclude that a typical Portlander would rate this restaurant at 3.5 stars?
    - What sources of bias are present in this sample?
- Suppose a year later, the restaurant still has 3.5 stars, but now with 1000 reviews. Does the verdic change?

## Sampling Example

- Suppose we want to know how Portlanders feel about a new Thai restaurant in Woodstock.
- The particular restaurant has a yelp rating of 3.5 stars with 100 reviews.
    - Can we conclude that a typical Portlander would rate this restaurant at 3.5 stars?
    - What sources of bias are present in this sample?
- Suppose a year later, the restaurant still has 3.5 stars, but now with 1000 reviews. Does the verdic change?
- Suppose a second Thai restaurant opens up nearby, with a yelp rating of 4 stars with 1000 reviews. Can we conclude Portlanders prefer the second restaurant to the first?

# Explanatory and Response Variables

- Consider the following question:

## Explanatory and Response Variables

- Consider the following question:
    - Is gross spending on healthcare higher or lower in countries with longer life expectancies?

# Explanatory and Response Variables

- Consider the following question:
    - Is gross spending on healthcare higher or lower in countries with longer life expectancies?
- If we suspect healthcare spending may affect life expectancy, then the former is the **explanatory variable**, while the latter is the **response variable**.

# Explanatory and Response Variables

- Consider the following question:
  - Is gross spending on healthcare higher or lower in countries with longer life expectancies?
- If we suspect healthcare spending may affect life expectancy, then the former is the **explanatory variable**, while the latter is the **response variable**.
  - Ultimately, these terms are labels to help keep track of which variables we suspect may affect the others. There may turn out to be no actual causal link between the two (or the link may be the reverse of what we suspect)

# Explanatory and Response Variables

- Consider the following question:
  - Is gross spending on healthcare higher or lower in countries with longer life expectancies?
- If we suspect healthcare spending may affect life expectancy, then the former is the **explanatory variable**, while the latter is the **response variable**.
  - Ultimately, these terms are labels to help keep track of which variables we suspect may affect the others. There may turn out to be no actual causal link between the two (or the link may be the reverse of what we suspect)
- Two types of data collection methods:
  1. Observational studies, where researchers do not interfere with how data arises.
  2. Random experiment, where individuals are assigned to group and a random treatment is assigned.

## Explanatory and Response Variables

- Consider the following question:
    - Is gross spending on healthcare higher or lower in countries with longer life expectancies?
- If we suspect healthcare spending may affect life expectancy, then the former is the **explanatory variable**, while the latter is the **response variable**.
    - Ultimately, these terms are labels to help keep track of which variables we suspect may affect the others. There may turn out to be no actual causal link between the two (or the link may be the reverse of what we suspect)
- Two types of data collection methods:
    1. Observational studies, where researchers do not interfere with how data arises.
    2. Random experiment, where individuals are assigned to group and a random treatment is assigned.
- Usually, only random experiments may allow researchers to conclude a causal link between explanatory and response variables.

Section 2

## Observational Studies

## Observational Studies and Association

- Generally, data in observational studies are collected only by monitoring what occurs. Thus, are only sufficient to show associations between variables.

## Observational Studies and Association

- Generally, data in observational studies are collected only by monitoring what occurs. Thus, are only sufficient to show associations between variables.
- An observational study tracks caffiene consumption in pregnant women and miscarriage rates, and finds the women who consumed more caffiene were more likely to have a miscarriange. Does this mean caffiene causes miscarriages?

# Observational Studies and Association

- Generally, data in observational studies are collected only by monitoring what occurs. Thus, are only sufficient to show associations between variables.
- An observational study tracks caffiene consumption in pregnant women and miscarriage rates, and finds the women who consumed more caffiene were more likely to have a miscarriange. Does this mean caffiene causes miscarriages?
    - Maybe? But further research actually suggests the presence of a **confounding variable**.

# Observational Studies and Association

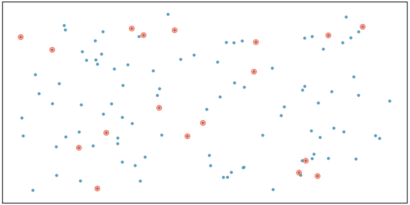
- Generally, data in observational studies are collected only by monitoring what occurs. Thus, are only sufficient to show associations between variables.
- An observational study tracks caffiene consumption in pregnant women and miscarriage rates, and finds the women who consumed more caffiene were more likely to have a miscarriange. Does this mean caffiene causes miscarriages?
    - Maybe? But further research actually suggests the presence of a **confounding variable**.
    - Women with healthy pregnancies have a higher rate of nausea during the 1st trimester than those with an unhealthy pregnancy, which often inhibits desire for coffee/tea in individuals who previously consumed these beverages.

# Observational Studies and Association

- Generally, data in observational studies are collected only by monitoring what occurs. Thus, are only sufficient to show associations between variables.
- An observational study tracks caffiene consumption in pregnant women and miscarriage rates, and finds the women who consumed more caffiene were more likely to have a miscarriange. Does this mean caffiene causes miscarriages?
  - Maybe? But further research actually suggests the presence of a **confounding variable**.
  - Women with healthy pregnancies have a higher rate of nausea during the 1st trimester than those with an unhealthy pregnancy, which often inhibits desire for coffee/tea in individuals who previously consumed these beverages.
  - As a result, women with early pregnancy complications (leading to increased risk of miscarriage) are less likely to stop consuming caffiene.

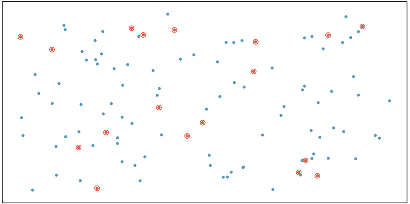
# Sampling Methods (SRS)

- **SRS**: Randomly select individuals from the population so that each individual has equal likelihood of being selected.



## Sampling Methods (SRS)

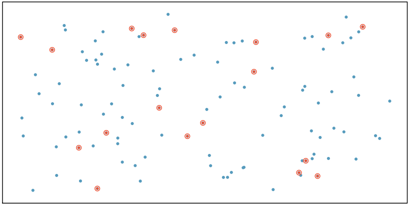
- **SRS**: Randomly select individuals from the population so that each individual has equal likelihood of being selected.



- Advantages:
  - Typically provides better representation compared to larger, non-random samples
  - Relatively simple to implement and analyze
  - Non-biased
  - Provides effective theoretical baseline

# Sampling Methods (SRS)

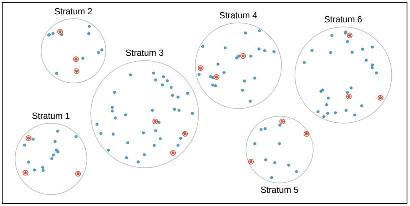
- **SRS**: Randomly select individuals from the population so that each individual has equal likelihood of being selected.



- Advantages:
  - Typically provides better representation compared to larger, non-random samples
  - Relatively simple to implement and analyze
  - Non-biased
  - Provides effective theoretical baseline
- Disadvantages:
  - May not be as precise as other sampling techniques
  - Can be difficult to perform in practice

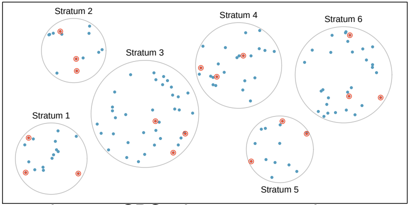
# Sampling Methods (Stratified)

- **Stratified**: Strata are made up of similar individuals, then an SRS is taken from each stratum.



# Sampling Methods (Stratified)

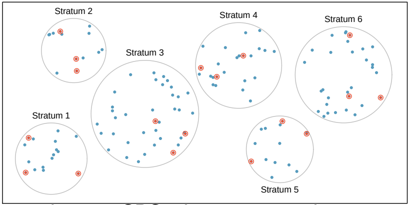
- **Stratified**: Strata are made up of similar individuals, then an SRS is taken from each stratum.



- Advantages:
  - Can be more precise than an SRS, thus requiring lower sample size
  - Hedges against non-representative samples
  - Strata proportions can be adjusted to ensure sufficient data to support analysis

## Sampling Methods (Stratified)

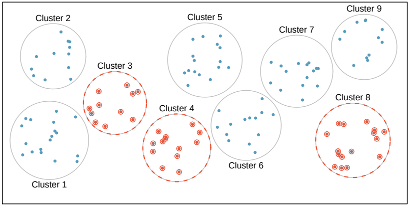
- **Stratified**: Strata are made up of similar individuals, then an SRS is taken from each stratum.



- Advantages:
  - Can be more precise than an SRS, thus requiring lower sample size
  - Hedges against non-representative samples
  - Strata proportions can be adjusted to ensure sufficient data to support analysis
- Disadvantages:
  - Requires more administrative labor in implementation
  - Statistical analysis is more complex
  - Cannot always be implemented

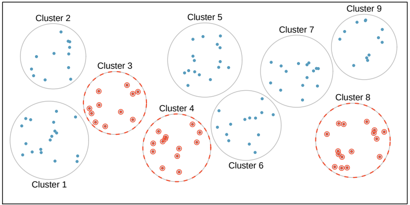
# Sampling Methods (Clustered)

- **Clustered**: An SRS is taken of non-homogeneous clusters. A sample is formed from all observations in those clusters.



# Sampling Methods (Clustered)

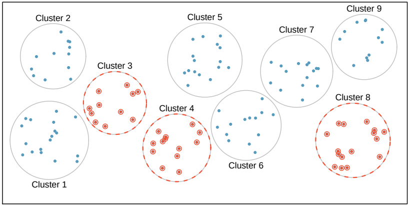
- **Clustered**: An SRS is taken of non-homogeneous clusters. A sample is formed from all observations in those clusters.



- Advantages:
  - Can be used when it is difficult or impossible to create complete list of population
  - Useful when population is naturally concentrated in heterogeneous groups
  - Often more cost/time effective per sample size than alternatives

# Sampling Methods (Clustered)

- **Clustered**: An SRS is taken of non-homogeneous clusters. A sample is formed from all observations in those clusters.



- Advantages:
    - Can be used when it is difficult or impossible to create complete list of population
    - Useful when population is naturally concentrated in heterogeneous groups
    - Often more cost/time effective per sample size than alternatives
- Disadvantages:
    - Is less precise than simple or statified sampling
    - Statistical analysis is more complex
    - Cannot always be implemented

Section 3

## Experiments

## Principles of Experiment Design

- Modern randomized experiments are built on 4 principles:

## Principles of Experiment Design

- Modern randomized experiments are built on 4 principles:
    1. **Controlling**: Treatments of interest are compared to a control group receiving no treatment.

## Principles of Experiment Design

- Modern randomized experiments are built on 4 principles:

  1. **Controlling**: Treatments of interest are compared to a control group receiving no treatment.
  2. **Randomized**: Patients are sorted into treatment groups randomly to account for variables that cannot be controlled.

# Principles of Experiment Design

- Modern randomized experiments are built on 4 principles:
    1. **Controlling**: Treatments of interest are compared to a control group receiving no treatment.
    2. **Randomized**: Patients are sorted into treatment groups randomly to account for variables that cannot be controlled.
    3. **Replicable**: Methodology should be thoroughly documented so that later researchers can replicate study to verify findings.

# Principles of Experiment Design

- Modern randomized experiments are built on 4 principles:
    1. **Controlling**: Treatments of interest are compared to a control group receiving no treatment.
    2. **Randomized**: Patients are sorted into treatment groups randomly to account for variables that cannot be controlled.
    3. **Replicable**: Methodology should be thoroughly documented so that later researchers can replicate study to verify findings.
    4. **Blocking**: If variables are suspected to affect response variable, subjects are first grouped into blocks based on these variables.

## Blocking Example

- Suppose we would like to design an experiment to investigate if a diet high in nitrates improves lung function.

## Blocking Example

- Suppose we would like to design an experiment to investigate if a diet high in nitrates improves lung function.
    - Explanatory variable: nitrate content of diet.
    - Response variable: exhaustion measured by O2 saturation.

## Blocking Example

- Suppose we would like to design an experiment to investigate if a diet high in nitrates improves lung function.
    - Explanatory variable: nitrate content of diet.
    - Response variable: exhaustion measured by O2 saturation.

    - Treatment: nitrate dietary supplement (powdered beet)
    - Control: No supplement

# Blocking Example

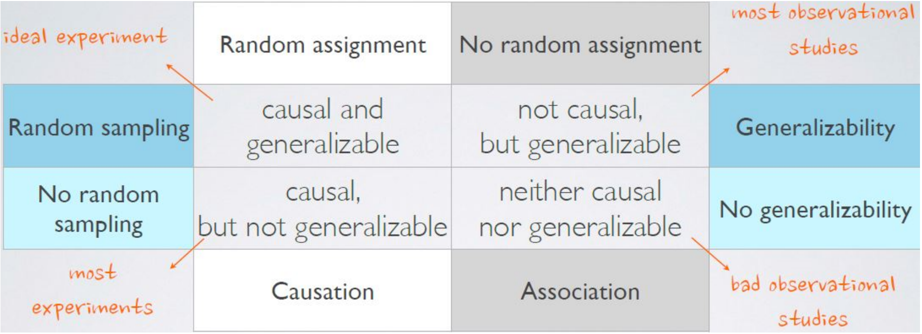
- Suppose we would like to design an experiment to investigate if a diet high in nitrates improves lung function.
    - Explanatory variable: nitrate content of diet.
    - Response variable: exhaustion measured by O2 saturation.
    - Treatment: nitrate dietary supplement (powdered beet)
    - Control: No supplement
- It is suspected that nitrate supplements may effect professional and amateur athletes differently, and so subjects are blocked for pro status:

# Blocking Example

- Suppose we would like to design an experiment to investigate if a diet high in nitrates improves lung function.
  - Explanatory variable: nitrate content of diet.
  - Response variable: exhaustion measured by O2 saturation.

  - Treatment: nitrate dietary supplement (powdered beet)
  - Control: No supplement
- It is suspected that nitrate supplements may effect professional and amateur athletes differently, and so subjects are blocked for pro status:
  1. Divide SRS into pro and amateur blocks.
  2. Randomly assign pro athletes to treatment and control groups.
  3. Similarly, randomly assign amateur athletes to treatment and control groups.
  4. Ensure pro/amateur status is equally represented in treatment and control groups.

# Random Sampling vs. Random Assignment



| ideal experiment | Random assignment | No random assignment | most observational studies |
|---|---|---|---|
| Random sampling | causal and generalizable | not causal, but generalizable | Generalizability |
| No random sampling | causal, but not generalizable | neither causal nor generalizable | No generalizability |
| most experiments | Causation | Association | bad observational studies |