Data Collection

Nate Wells

Math 141, 2/11/22

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 together. 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 together. 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 together. 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 together. 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 under-representation 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

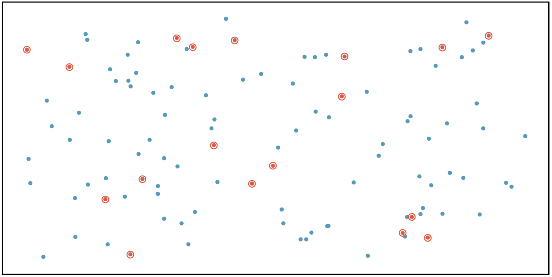
- Suppose we want to know how Portlanders feel about a new Thai restaurant in Woodstock.
- The particular restaurant has a yelp rating of 3.5 stars with 100 reviews.
  - Can we conclude that a typical Portlander would rate this restaurant at 3.5 stars?
  - What sources of bias are present in this sample?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

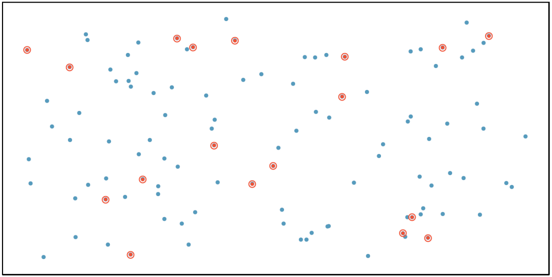
Sampling Methods (SRS)

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



Sampling Methods (SRS)

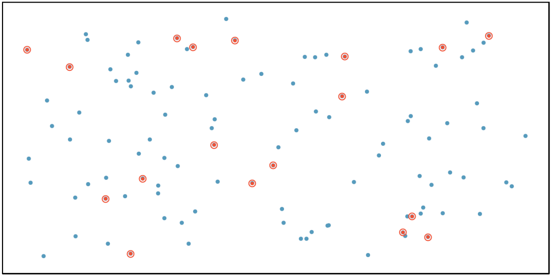
- **SRS**: Randomly select individuals from the population so that each individual has equal likelihood of being selected.



- 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)

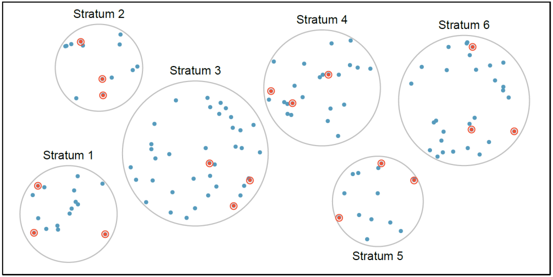
- **SRS**: Randomly select individuals from the population so that each individual has equal likelihood of being selected.



- Disadvantages:
  - May not be as precise as other sampling techniques
  - Can be difficult to perform in practice

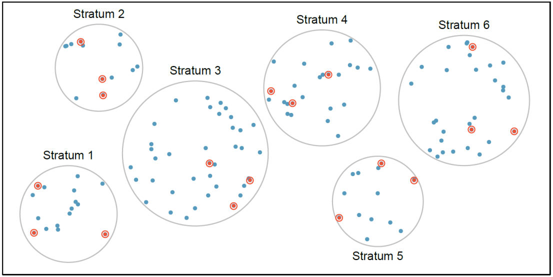
Sampling Methods (Stratified)

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



Sampling Methods (Stratified)

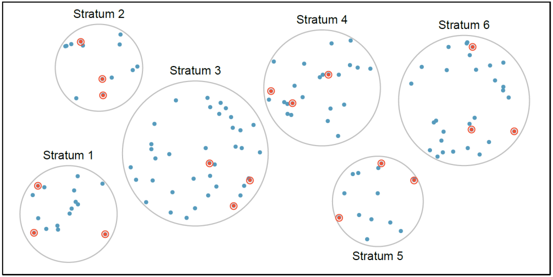
- **Stratified**: Strata are made up of similar individuals, then an SRS is taken from each stratum.



- 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)

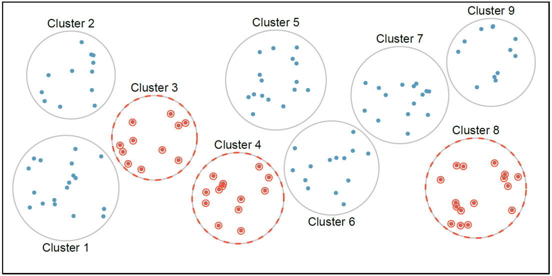
- **Stratified**: Strata are made up of similar individuals, then an SRS is taken from each stratum.



- Disadvantages:
  - Requires more administrative labor in implementation
  - Statistical analysis is more complex
  - Cannot always be implemented

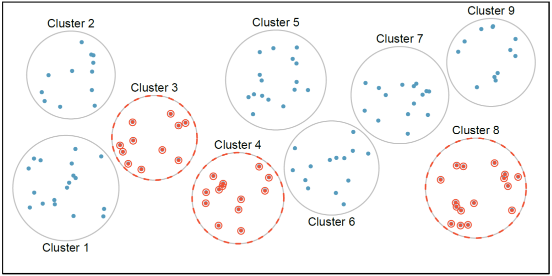
Sampling Methods (Clustered)

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



Sampling Methods (Clustered)

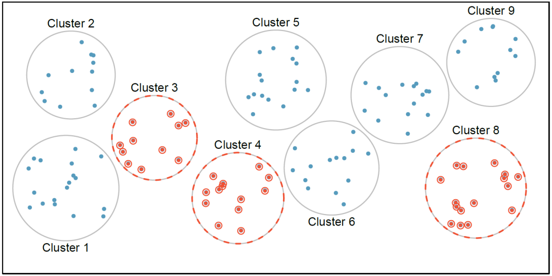
- **Clustered**: An SRS is taken of non-homogeneous clusters. A sample is formed from all observations in those clusters.



- 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.



- Disadvantages:
  - Is less precise than simple or stratified sampling
  - Statistical analysis is more complex
  - Cannot always be implemented

Section 2

Assessing Relationships Between Variables

Explanatory and Response Variables

- Consider the following question:

Explanatory and Response Variables

- Consider the following question:
  - Is total spending on health care higher or lower in countries with longer life expectancy?

Explanatory and Response Variables

- Consider the following question:
    - Is total spending on health care higher or lower in countries with longer life expectancy?
- If we suspect health care 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 total spending on health care higher or lower in countries with longer life expectancy?
- If we suspect health care 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 total spending on health care higher or lower in countries with longer life expectancy?
- If we suspect health care 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 total spending on health care higher or lower in countries with longer life expectancy?
- If we suspect health care 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.

Correlation and Causation

- Two quantitative (or ordinal categorical) variables are **correlated** if increasing values of one variable are coupled with consistently increasing or decreasing values of the other.

Correlation and Causation

- Two quantitative (or ordinal categorical) variables are **correlated** if increasing values of one variable are coupled with consistently increasing or decreasing values of the other.
- Two variables are **causally linked** if changing the value of one variable actively produces a change in the other variable.

Correlation and Causation

- Two quantitative (or ordinal categorical) variables are **correlated** if increasing values of one variable are coupled with consistently increasing or decreasing values of the other.
- Two variables are **causally linked** if changing the value of one variable actively produces a change in the other variable.
- Correlation is bi-directional: If X is correlated with Y, then Y is necessarily correlated with X

Correlation and Causation

- Two quantitative (or ordinal categorical) variables are **correlated** if increasing values of one variable are coupled with consistently increasing or decreasing values of the other.
- Two variables are **causally linked** if changing the value of one variable actively produces a change in the other variable.
- Correlation is bi-directional: If X is correlated with Y, then Y is necessarily correlated with X
- Causality can be mono-directional: It is possible for changes in X to cause changes in Y, but for changes in Y *not* to cause changes in X.

Correlation and Causation

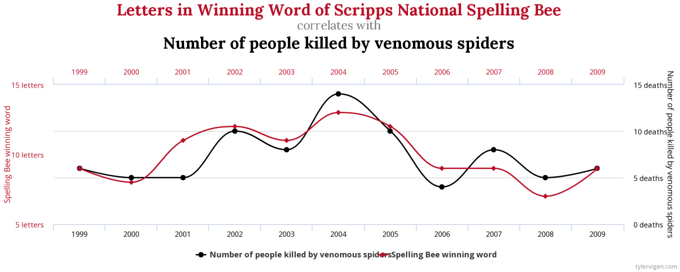
- Two quantitative (or ordinal categorical) variables are **correlated** if increasing values of one variable are coupled with consistently increasing or decreasing values of the other.
- Two variables are **causally linked** if changing the value of one variable actively produces a change in the other variable.
- Correlation is bi-directional: If X is correlated with Y, then Y is necessarily correlated with X
- Causality can be mono-directional: It is possible for changes in X to cause changes in Y, but for changes in Y *not* to cause changes in X.
- If variables X and Y are correlated, there are 4 possible explanations:
  1. Changes in X cause changes in Y
  2. Changes in Y cause changes in X
  3. Changes in a third variable Z cause changes in *both* X and Y
  4. The observed correlation in X and Y is due to chance.

Correlation Due to Chance

- **The Problem of Multiple Comparisons**: Given enough variables, it is improbable not to observe a correlation between at least two of them.

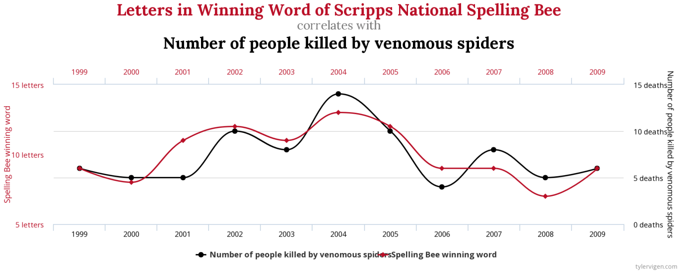
Correlation Due to Chance

- **The Problem of Multiple Comparisons**: Given enough variables, it is improbable not to observe a correlation between at least two of them.



Correlation Due to Chance

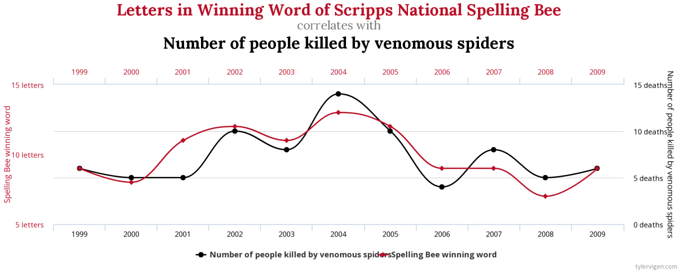
- **The Problem of Multiple Comparisons**: Given enough variables, it is improbable not to observe a correlation between at least two of them.



- How do we rule out spurious correlations?

Correlation Due to Chance

- **The Problem of Multiple Comparisons**: Given enough variables, it is improbable not to observe a correlation between at least two of them.



- How do we rule out spurious correlations?
  - Gather more data. If the correlation occurred by chance just due to sampling, the relationship is unlikely to be repeated in an independent sample.

Confounding Variables

- Two variables may be correlated if both are causally linked to a third variable.

Confounding Variables

- Two variables may be correlated if both are causally linked to a third variable.

  In Spring 2021, Public Health England published a report investigating the link between COVID-19 variants of concern, vaccination, and negative health outcomes.

Confounding Variables

- Two variables may be correlated if both are causally linked to a third variable.

  In Spring 2021, Public Health England published a report investigating the link between COVID-19 variants of concern, vaccination, and negative health outcomes.

  The report indicates that among the vaccinated population, mortality rate due to Delta was 0.41%, while among the unvaccinated population, mortality rate was 0.17%

Confounding Variables

- Two variables may be correlated if both are causally linked to a third variable.

  In Spring 2021, Public Health England published a report investigating the link between COVID-19 variants of concern, vaccination, and negative health outcomes.

  The report indicates that among the vaccinated population, mortality rate due to Delta was 0.41%, while among the unvaccinated population, mortality rate was 0.17%

- Does this indicate that vaccination actually *increases* mortality?

Confounding Variables

- Two variables may be correlated if both are causally linked to a third variable.

  In Spring 2021, Public Health England published a report investigating the link between COVID-19 variants of concern, vaccination, and negative health outcomes.

  The report indicates that among the vaccinated population, mortality rate due to Delta was 0.41%, while among the unvaccinated population, mortality rate was 0.17%

- Does this indicate that vaccination actually *increases* mortality?
  - In early 2021, vaccination rates were significantly higher among individuals 50 and over.
  - Mortality due to respiratory infection is also significantly higher in these individuals.
  - Breaking down mortality rates by *age* shows that in each age group, mortality was higher among unvaccinated than vaccinated individuals

Confounding Variables

- Two variables may be correlated if both are causally linked to a third variable.

  In Spring 2021, Public Health England published a report investigating the link between COVID-19 variants of concern, vaccination, and negative health outcomes.

  The report indicates that among the vaccinated population, mortality rate due to Delta was 0.41%, while among the unvaccinated population, mortality rate was 0.17%

- Does this indicate that vaccination actually *increases* mortality?
  - In early 2021, vaccination rates were significantly higher among individuals 50 and over.
  - Mortality due to respiratory infection is also significantly higher in these individuals.
  - Breaking down mortality rates by *age* shows that in each age group, mortality was higher among unvaccinated than vaccinated individuals
- How do we rule out confounding variables?

Confounding Variables

- Two variables may be correlated if both are causally linked to a third variable.

  In Spring 2021, Public Health England published a report investigating the link between COVID-19 variants of concern, vaccination, and negative health outcomes.

  The report indicates that among the vaccinated population, mortality rate due to Delta was 0.41%, while among the unvaccinated population, mortality rate was 0.17%

- Does this indicate that vaccination actually *increases* mortality?
  - In early 2021, vaccination rates were significantly higher among individuals 50 and over.
  - Mortality due to respiratory infection is also significantly higher in these individuals.
  - Breaking down mortality rates by *age* shows that in each age group, mortality was higher among unvaccinated than vaccinated individuals
- How do we rule out confounding variables?
  - Create models that include possible confounding variables

Confounding Variables

- Two variables may be correlated if both are causally linked to a third variable.

  In Spring 2021, Public Health England published a report investigating the link between COVID-19 variants of concern, vaccination, and negative health outcomes.

  The report indicates that among the vaccinated population, mortality rate due to Delta was 0.41%, while among the unvaccinated population, mortality rate was 0.17%

- Does this indicate that vaccination actually *increases* mortality?
  - In early 2021, vaccination rates were significantly higher among individuals 50 and over.
  - Mortality due to respiratory infection is also significantly higher in these individuals.
  - Breaking down mortality rates by *age* shows that in each age group, mortality was higher among unvaccinated than vaccinated individuals
- How do we rule out confounding variables?
  - Create models that include possible confounding variables
  - Design experiments that control for possible confounding variables

Reverse Causality

- If two variables are causally linked, correlation alone will not indicate which is the cause of the other.

Reverse Causality

- If two variables are causally linked, correlation alone will not indicate which is the cause of the other.

  Several scientific studies during the 1950s and 1960s demonstrate that infants who receive prolonged and exclusive breastfeeding grow more slowly during the first year of life than those who do not.

Reverse Causality

- If two variables are causally linked, correlation alone will not indicate which is the cause of the other.

  Several scientific studies during the 1950s and 1960s demonstrate that infants who receive prolonged and exclusive breastfeeding grow more slowly during the first year of life than those who do not.

Reverse Causality

- If two variables are causally linked, correlation alone will not indicate which is the cause of the other.

  Several scientific studies during the 1950s and 1960s demonstrate that infants who receive prolonged and exclusive breastfeeding grow more slowly during the first year of life than those who do not.

- Does breastfeeding cause reduced infant growth?

Reverse Causality

- If two variables are causally linked, correlation alone will not indicate which is the cause of the other.

  Several scientific studies during the 1950s and 1960s demonstrate that infants who receive prolonged and exclusive breastfeeding grow more slowly during the first year of life than those who do not.

- Does breastfeeding cause reduced infant growth?
  - Perhaps not. A randomized experiment involving 17,000 Belarusian infants between 1996 and 1997 found that smaller size was strongly associated with subsequent weaning and discontinuation of exclusive breastfeeding in each follow-up interval (even after adjusting for confounding variables.)

Reverse Causality

- If two variables are causally linked, correlation alone will not indicate which is the cause of the other.

  Several scientific studies during the 1950s and 1960s demonstrate that infants who receive prolonged and exclusive breastfeeding grow more slowly during the first year of life than those who do not.

- Does breastfeeding cause reduced infant growth?
  - Perhaps not. A randomized experiment involving 17,000 Belarusian infants between 1996 and 1997 found that smaller size was strongly associated with subsequent weaning and discontinuation of exclusive breastfeeding in each follow-up interval (even after adjusting for confounding variables.)
- How do you rule out reverse causation?

Reverse Causality

- If two variables are causally linked, correlation alone will not indicate which is the cause of the other.

  Several scientific studies during the 1950s and 1960s demonstrate that infants who receive prolonged and exclusive breastfeeding grow more slowly during the first year of life than those who do not.

- Does breastfeeding cause reduced infant growth?
  - Perhaps not. A randomized experiment involving 17,000 Belarusian infants between 1996 and 1997 found that smaller size was strongly associated with subsequent weaning and discontinuation of exclusive breastfeeding in each follow-up interval (even after adjusting for confounding variables.)
- How do you rule out reverse causation?
  - Investigate the temporal order of events.
  - Design an experiment where theorized cause is administered as treatment.

Correlation and Causation

- Correlation does not imply causation. But it also does not imply not causation.

Correlation and Causation

- Correlation does not imply causation. But it also does not imply not causation.

  In 1950, Hill and Doll published a controlled study showing extremely strong association between smoking and lung cancer.

Correlation and Causation

- Correlation does not imply causation. But it also does not imply not causation.

  In 1950, Hill and Doll published a controlled study showing extremely strong association between smoking and lung cancer.

  In a 1958 article in Nature, (in)famous statistician R. A. Fisher presented a case that smoking ***does not*** *cause lung cancer, arguing that:*

  "If, for example, it were possible to infer that smoking cigarettes is a cause of this disease, it would equally be possible to infer on exactly similar grounds that inhaling cigarette smoke was a practice of considerable prophylactic value in preventing the disease, for the practice of inhaling is rarer among patients with cancer of the lung than with others."

Correlation and Causation

- Correlation does not imply causation. But it also does not imply not causation.

  In 1950, Hill and Doll published a controlled study showing extremely strong association between smoking and lung cancer.

  In a 1958 article in Nature, (in)famous statistician R. A. Fisher presented a case that smoking ***does not*** *cause lung cancer, arguing that:*

  "If, for example, it were possible to infer that smoking cigarettes is a cause of this disease, it would equally be possible to infer on exactly similar grounds that inhaling cigarette smoke was a practice of considerable prophylactic value in preventing the disease, for the practice of inhaling is rarer among patients with cancer of the lung than with others."

  That is, according to Fisher, what if people disposed to cancer turn to cigarettes to relieve discomfort?

- Fisher did not disagree with the statistical analysis that smoking and cancer were highly correlated.

Correlation and Causation

- Correlation does not imply causation. But it also does not imply not causation.

  In 1950, Hill and Doll published a controlled study showing extremely strong association between smoking and lung cancer.

  *In a 1958 article in Nature, (in)famous statistician R. A. Fisher presented a case that smoking **does not** cause lung cancer, arguing that:*

  "If, for example, it were possible to infer that smoking cigarettes is a cause of this disease, it would equally be possible to infer on exactly similar grounds that inhaling cigarette smoke was a practice of considerable prophylactic value in preventing the disease, for the practice of inhaling is rarer among patients with cancer of the lung than with others."

  That is, according to Fisher, what if people disposed to cancer turn to cigarettes to relieve discomfort?

- Fisher did not disagree with the statistical analysis that smoking and cancer were highly correlated.
- So how do we know that Fisher was wrong? (He was)

Hill's Criteria for Causation

- In 1965, Austin Bradford Hill outlined 9 criteria for inferring causality:

Hill's Criteria for Causation

- In 1965, Austin Bradford Hill outlined 9 criteria for inferring causality:
  1. **Strength** Causal events should have strong correlation.
  2. **Consistency** Different studies should show similar effect.
  3. **Specificity** A single cause should lead to a single effect.
  4. **Temporality** The effect should occur before the cause.
  5. **Gradient** Greater exposure to cause should correspond to greater size of effect
  6. **Plausibility** A plausible mechanism should exist linking cause and effect.
  7. **Coherence** A cause and effect relationship should not conflict with other known relationships
  8. **Experimental Evidence** A cause and effect relationship should be evident in randomized experiment.
  9. **Analogy** A cause and effect relationship should also be observed in other similar phenomena

Hill's Criteria for Causation

- In 1965, Austin Bradford Hill outlined 9 criteria for inferring causality:
  1. **Strength** Causal events should have strong correlation.
  2. **Consistency** Different studies should show similar effect.
  3. **Specificity** A single cause should lead to a single effect.
  4. **Temporality** The effect should occur before the cause.
  5. **Gradient** Greater exposure to cause should correspond to greater size of effect
  6. **Plausibility** A plausible mechanism should exist linking cause and effect.
  7. **Coherence** A cause and effect relationship should not conflict with other known relationships
  8. **Experimental Evidence** A cause and effect relationship should be evident in randomized experiment.
  9. **Analogy** A cause and effect relationship should also be observed in other similar phenomena
- Are these *absolutely* necessary to prove causality?

Hill's Criteria for Causation

- In 1965, Austin Bradford Hill outlined 9 criteria for inferring causality:
  1. **Strength** Causal events should have strong correlation.
  2. **Consistency** Different studies should show similar effect.
  3. **Specificity** A single cause should lead to a single effect.
  4. **Temporality** The effect should occur before the cause.
  5. **Gradient** Greater exposure to cause should correspond to greater size of effect
  6. **Plausibility** A plausible mechanism should exist linking cause and effect.
  7. **Coherence** A cause and effect relationship should not conflict with other known relationships
  8. **Experimental Evidence** A cause and effect relationship should be evident in randomized experiment.
  9. **Analogy** A cause and effect relationship should also be observed in other similar phenomena
- Are these *absolutely* necessary to prove causality?
  - No. But they are good guidelines.

Section 3

Experiments and Observational Studies

Principles of Experiment Design

- The randomized experiment is the standard tool used to demonstrate causal relationship between variables.

Principles of Experiment Design

- The randomized experiment is the standard tool used to demonstrate causal relationship between variables.
- Modern randomized experiments are built on 4 principles:

Principles of Experiment Design

- The randomized experiment is the standard tool used to demonstrate causal relationship between variables.
- 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

- The randomized experiment is the standard tool used to demonstrate causal relationship between variables.
- 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

- The randomized experiment is the standard tool used to demonstrate causal relationship between variables.
- 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

- The randomized experiment is the standard tool used to demonstrate causal relationship between variables.
- 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.

Observational Studies and Association

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

Observational Studies and Association

- Generally, data in observational studies are collected only by monitoring what occurs. Thus, they are usually only sufficient to show associations between variables.
- So why conduct observational studies at all?

Observational Studies and Association

- Generally, data in observational studies are collected only by monitoring what occurs. Thus, they are usually only sufficient to show associations between variables.
- So why conduct observational studies at all?
  - Sometimes, observational studies are the *only* tool available for assessing the state of the world in order to make decisions.

Observational Studies and Association

- Generally, data in observational studies are collected only by monitoring what occurs. Thus, they are usually only sufficient to show associations between variables.
- So why conduct observational studies at all?
    - Sometimes, observational studies are the *only* tool available for assessing the state of the world in order to make decisions.
- Experiments may be unethical
    - To study whether smoking causes cancer, we cannot randomly force participants to smoke or to not smoke.

Observational Studies and Association

- Generally, data in observational studies are collected only by monitoring what occurs. Thus, they are usually only sufficient to show associations between variables.
- So why conduct observational studies at all?
    - Sometimes, observational studies are the *only* tool available for assessing the state of the world in order to make decisions.
- Experiments may be unethical
    - To study whether smoking causes cancer, we cannot randomly force participants to smoke or to not smoke.
- Experiments may be temporally impossible
    - To study whether gender influenced survival rate on the Titanic, we only have historical records to look at.

Observational Studies and Association

- Generally, data in observational studies are collected only by monitoring what occurs. Thus, they are usually only sufficient to show associations between variables.
- So why conduct observational studies at all?
  - Sometimes, observational studies are the *only* tool available for assessing the state of the world in order to make decisions.
- Experiments may be unethical
  - To study whether smoking causes cancer, we cannot randomly force participants to smoke or to not smoke.
- Experiments may be temporally impossible
  - To study whether gender influenced survival rate on the Titanic, we only have historical records to look at.
- Experiments may not be manufacturable
  - To study whether high unemployment rate leads to presidential losses for the incumbent party, we cannot create new presidential races.

Observational Studies and Association

- Generally, data in observational studies are collected only by monitoring what occurs. Thus, they are usually only sufficient to show associations between variables.
- So why conduct observational studies at all?
  - Sometimes, observational studies are the *only* tool available for assessing the state of the world in order to make decisions.
- Experiments may be unethical
  - To study whether smoking causes cancer, we cannot randomly force participants to smoke or to not smoke.
- Experiments may be temporally impossible
  - To study whether gender influenced survival rate on the Titanic, we only have historical records to look at.
- Experiments may not be manufacturable
  - To study whether high unemployment rate leads to presidential losses for the incumbent party, we cannot create new presidential races.
- Experiments of appropriate size may be prohibitively expensive
  - Experiments of small or moderate size often include uncontrolled confounding variables

Random Sampling vs. Random Assignment

- Statistical investigations can incorporate two sources of randomization:

Random Sampling vs. Random Assignment

- Statistical investigations can incorporate two sources of randomization:

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