Classifying Data

Two Types of Data

Numerical data (Quantitative data)

- Tell us how much or how many.
  - Discrete
  - Continuous
- Example: number of siblings, weight, temperature, …

Categorical data (Qualitative data)

- Tell us what type or what kind.
  - Nominal
  - Ordinal
- Example: eye color, zip code, major in school, …

Variable Coding

Caution!

It is not always obvious if a variable is numeric or categorical just by whether the values are numerical or not. It is important to consider what the values represent in context.

Categorical variables can be coded as numerical values

- Area codes (e.g. 310, 626, 800)
- Weekday / Weekend → 1 / 0
- Yes / No → 1 / 0

Numerical variables can be coded into categories.

- Income 10K, 30K, 60K, 90K … → low, middle, high
- Age 1m, 3m, 9m, 20m, … → newborn, infant, toddler, etc

Numerical or Categorical?

Name	Male	Age in years	Height	Hair Color	Zip code
Leslie	1	34	62.2	Blonde	90001
Ben	0	35	70.0	Brown	90011
Ron	1	49	71.3	Brown	90014
April	0	20	66.5	Black	90025
Andy	1	28	74.1	Brown	90035

Organizing Categorical Data

Frequency Tables

- **Frequencies (or counts):** the number of times a value is observed in a data set.
- **Relative frequency:** the proportion/percentage of times a value is observed in a data set.

Eye color	Frequency	Relative frequency
Brown	30	30/100 = 30%
Blue	20	20/100 = 20%
Black	25	25/100 = 25%
Green	15	15/100 = 15%
Yellow	10	10/100 = 10%

Summarize Two Categorical Variables

Example: A youth Behavior Risk Survey was conducted to study the relationship between gender and whether the respondent always wears a seat belt when riding in or driving a car.

Male	Not Always
1	1
1	1
1	0
1	0
1	0
0	1
0	1
0	1
0	0
0	0
0	0
0	0
0	0
0	0
0	0

1. How many observations are in the sample?
2. How many people do not always wear seat belts?
3. How many of those who always wear seat belts are males?
4. What percent always wear seat belts?
5. Are males in the sample more likely to take the risk of not wearing a seat belt?
6. Being a man results in not wearing seat belts.

   Yes or No?

Two-way Tables

Summarize two categorical variables, display frequency of combinations of categories.

1. How many observations are in the sample?
2. How many people do not always wear seat belts?
3. How many of those who always wear seat belts are males?
4. What percent always wear seat belts?
5. Are males in the sample more likely to take the risk of not wearing a seat belt?
6. Being a man results in not wearing seat belts.

   Yes or No?

Male

Not Always	1	0	Total
1	2	3	5
0	3	7	10
Total	5	10	15

Two-way Tables

Summarize two categorical variables, display frequency of combinations of categories.

1. How many observations are in the sample?
2. How many people do not always wear seat belts?
3. How many of those who always wear seat belts are males?
4. What percent always wear seat belts?
5. Are males in the sample more likely to take the risk of not wearing a seat belt?
6. Being a man results in not wearing seat belts.

   Yes or No?

Male

Not Always	1	0	Total
1	2	3	5
0	3	7	10
Total	5	10	15

1. 15
2. 5
3. 3
4. 2/3, 66.7%
5. 40% > 30%
   Yes
6. NO!

Comparing Data

Data summary: number of sports-related injuries that were treated in U.S emergency rooms in 2009

Which team sports is the most dangerous?

Sport	Injuries
Baseball	165,842
Basketball	501,251
Bowling	20,878
Football	451,961
Ice hockey	19,035
Soccer	208,214
Softball	121,175
Tennis	23,611
Volleyball	60,159

Comparing Data

Data summary: number of sports-related injuries that were treated in U.S emergency rooms in 2009

Which team sports is the most dangerous?

Sport	Injuries	Participants
Baseball	165,842	11,500,000
Basketball	501,251	24,400,000
Bowling	20,878	45,000,000
Football	451,961	8,900,000
Ice hockey	19,035	3,100,000
Soccer	208,214	13,600,000
Softball	121,175	11,800,000
Tennis	23,611	10,800,000
Volleyball	60,159	10,700,000

Comparing Data

- The groups need to be similar.
- Percentages or rates are often better for comparisons.

Sport	Participants	Injuries	Rate of Injury per Participant	Rate of Injury per Thousand Participants
Baseball	11,500,000	165,842	0.01442	14.42
Basketball	24,400,000	501,251	0.02054	20.54
Bowling	45,000,000	20,878	0.00046	0.46
Football	8,900,000	451,961	0.05078	50.78
Ice hockey	3,100,000	19,035	0.00614	6.14
Soccer	13,600,000	208,214	0.01531	15.31
Softball	11,800,000	121,175	0.01027	10.27
Tennis	10,800,000	23,611	0.00219	2.19
Volleyball	10,700,000	60,159	0.00562	5.62

Collecting Data to Understand Causality

Collecting Data to Understand Causality

Often, the most important questions in science, business, and daily life are questions about causality.

"What if?" – Counterfactual Reasoning

- What if I did not take this medicine, will I still get better?
- What if the company uses YouTube for advertising, will it be better than TV commercials?
- What if I go to class, will my grade be higher?

Most questions about causality can be understood in terms of two variables:

- **Treatment variable** (predictor, independent/explanatory variable)
- **Outcome variable** (response/dependent variable)

Causality or Causation

Does the treatment variable cause changes in the outcome variable?

Cause ⟶ Effect

Association and Causation

Association: one variable provides information about another.

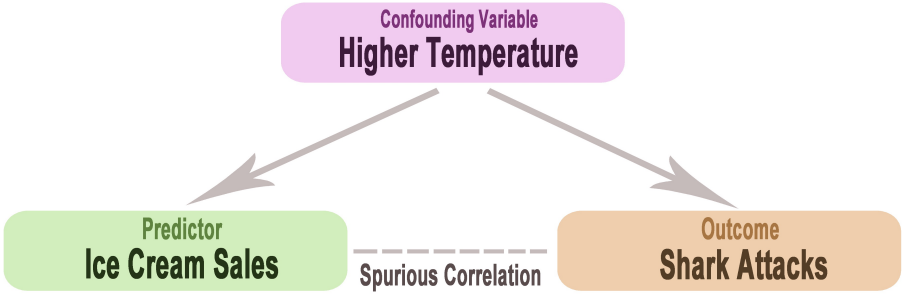
Two variables are associated if there is a relationship between them.

Caution! Association does NOT mean Causation!

Confounding Variable

A third variable that influences the variables of interest.

- Causes a difference between the two groups that could explain why the outcomes were different
- Causes the two variables of interest to falsely appear to be causal related.



An experiment that fails to take a confounding variable into account has poor internal validity.

Identify Potential Confounding Variable

Research question:

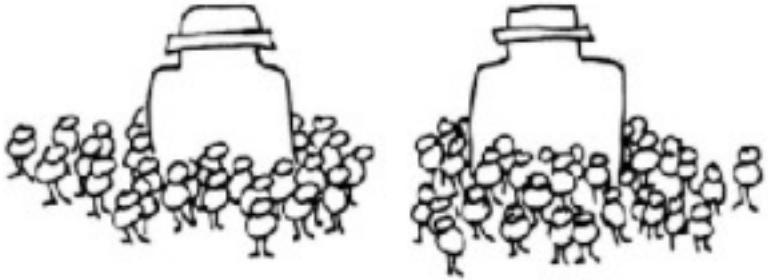
Do adults who prefer to drink beer, wine, and water differ in terms of their mean weights?

Data were collected from a sample of World Campus students to address the research question above. The researchers found that adults who preferred beer tended to weigh more than those who preferred wine.

Establishing Causality

We want to answer whether the treatment variable **causes** the changes in the outcome variable

- **Treatment group**: subjects who receive the treatment of interest
- **Control group**: subjects who do not receive the treatment



In order to conclude causality from a study, it is important to have both treatment and control groups, and for subjects in both groups to be identical in every way except for the treatment.

Producing Data

Two stages

1. **Sampling**
   - For the conclusions to be valid, we must have a **Representative sample** of the population.
   - A sample that is not representative is said to be **biased**.
   - One way to collect a sample that is representative is **Simple Random sampling** (**SRS**).
2. **Study design**
   - Studies should be designed to discover what we want to know about the variables of interest for the individuals in the sample.
   - Anecdotes (stories based one's experience) are not useful for making cause-and-effect conclusions about population.

Study Design

I. Observational study

- Subjects are not assigned to the treatment group or control group by the researchers.
- The subjects in the study are simply observed.
- Researcher do not interfere with the subjects other than taking measurements.

II. Sample survey

A particular type of observational study in which individuals report variables' values themselves, frequently by giving their opinions.

We need to be careful with cause-and-effect conclusion from observational studies alone because of potential confounding variables.

Study Design

III. Experimental Study

- The researcher actively manipulates the treatment variable.
- Subjects are assigned to the control/treatment group.
- At least one treatment variable to manipulate and one outcome variable to measure.
- The outcome variable is measured and compared between groups receiving different treatments.

It is possible to answer questions about population-level causal effects with experiments.

- Well-designed and well-executed controlled experiments are most important ways for answering questions about causality.
- Controlled experiment requires at least a treatment group and a control group.

Exercise

Suppose researchers want to determine whether people tend to snack more while they watch television. In other words, the researchers would like to explore the relationship between the explanatory variable "TV" (a categorical variable that takes the values "on"' and "not on") and the response variable "snack consumption."

Identify each of the following designs as being an observational study, a survey, or an experiment.

1. Recruit participants for a study. While they are presumably waiting to be interviewed, half of the individuals sit in a waiting room with snacks available and a TV on. The other half sit in a waiting room with snacks available and no TV, just magazines. Researchers determine whether people consume more snacks in the TV setting.

2. Recruit participants for a study. Give them journals to record hour by hour their activities the following day, including when they watch TV and when they consume snacks. Determine if snack consumption is higher during TV times.

3. Poll a sample of individuals with the following question: While watching TV, do you tend to snack: (a) less than usual; (b) more than usual; or (c) the same amount as usual?

Controlled Experiments

Key features:

1. Sample size
Large sample size ensures that we observe the full range of variability in the objects we study.

2. Random assignment
- Subjects are assigned to the control or treatment group by a **randomization procedure**
- Control for **confounding variables**.
- Balances out differences to make the groups **comparable**. Bias may occur for non-randomized assignment and the results are influenced in one particular direction.
- An experiment where subjects are randomly assigned is called a **randomized experiment**.

Controlled Experiments

Key features:

3. **Blinding --** Knowing what treatment was assigned can bias the study.
   - Researchers should be blind to the assignment
   - Participants should be blind to the assignment

When neither the researchers nor the participants know whether the participants are in the treatment or the comparison group, it is a **double-blinded** study.

The double-blind format helps prevent the bias that can result if one group acts differently from the other because they know they are being treated differently, or because the researchers treat the groups differently or evaluate them differently because of what the researchers hope or expect.

Controlled Experiments

Key features:

4. Placebos

- Anything that seems to be a "real" treatment,
  - e.g., harmless pill, a shot, or some other type of "fake" treatment.
- It is important that the comparison group receive attention similar to what the treatment group receives, so that both groups feel they are being treated the same by the researchers.
- This format controls for possible differences between groups that occur simply because some subjects are more likely than others to expect their treatment to be effective.

Causal Inference

Experimental study:

- In general, with a well-designed experiment we have a better chance of establishing causation than with an observational study.
- However, experiments are subject to certain pitfalls, and there are many situations in which an experiment is not an option.

Observational study:

- Because of the existence of a virtually unlimited number of potential confounding variables, we can never be 100% certain of a claim of causation based on an observational study.
- However, A well-designed observational study may still provide fairly convincing evidence of causation under the right circumstances.

Counterfactual causal inference is one the most important statistical ideas of the past 50 years!

(Paper by Andrew Gelman, Aki Vehtari, 2020)

Observational vs. Experimental

Does drinking coffee affect people's sleeping habit?

A study took random sample of adults and asked them about their coffee drinking and sleeping habits. The data showed that people who drank coffee daily were more likely to go to sleep later than those who didn't.

Identify:

- Treatment
- Outcome
- Observational study or experimental study?
- Can you conclude that drinking coffee causes people to sleep late?