Lecture 13: Introduction to Data

- Data Basics
- Observational Studies
- Designed Experiments
- Simpson's Paradox

Treating Chronic Fatigue Syndrome

- Objective: Evaluate the effectiveness of cognitive-behavior therapy for chronic fatigue syndrome.
- Participant pool: 142 patients who were recruited from referrals by primary care physicians and consultants to a hospital clinic specializing in chronic fatigue syndrome.
- Actual participants: Only 60 of the 142 referred patients entered the study. Some were excluded because they didn't meet the diagnostic criteria, some had other health issues, and some refused to be a part of the study.

Deale et. al. *Cognitive behavior therapy for chronic fatigue syndrome: A randomized controlled trial*. The American Journal of Psychiatry 154.3 (1997).

Study design

- Patients randomly assigned to treatment and control groups, 30 patients in each group:
  - *Treatment*: Cognitive behavior therapy – collaborative, educative, and with a behavioral emphasis. Patients were shown on how activity could be increased steadily and safely without exacerbating symptoms.
  - *Control:* Relaxation – No advice was given about how activity could be increased. Instead progressive muscle relaxation, visualization, and rapid relaxation skills were taught.

Results

The table below shows the distribution of patients with good outcomes at 6-month follow-up. Note that 7 patients dropped out of the study: 3 from the treatment and 4 from the control group.

		Good outcome		
		Yes	No	Total
Group	Treatment	19	8	27
	Control	5	21	26
	Total	24	29	53

- Proportion with good outcomes in treatment group:

$$19/27 \approx 0.70 \rightarrow 70\%$$

- Proportion with good outcomes in control group:

$$5/26 \approx 0.19 \rightarrow 19\%$$

Understanding the results

Do the data show a “real" difference between the groups?

- Suppose you flip a coin 100 times. While the chance a coin lands heads in any given coin flip is 50%, we probably won’t observe exactly 50 heads. This type of fluctuation is part of almost any type of data generating process.
- The observed difference between the two groups (70 - 19 = 51%) may be real, or may be due to natural variation.
- Since the difference is quite large, it is more believable that the difference is real.
- We need statistical tools to determine if the difference is so large that we should reject the notion that it was due to chance.

Generalizing the results

Are the results of this study generalizable to all patients with chronic fatigue syndrome?

These patients had specific characteristics and volunteered to be a part of this study, therefore they may not be representative of all patients with chronic fatigue syndrome. While we cannot immediately generalize the results to all patients, this first study is encouraging. The method works for patients with some narrow set of characteristics, and that gives hope that it will work, at least to some degree, with other patients.

Example: class survey

Students in an introductory statistics course were asked the following questions as part of a class survey:

1. What is your gender, male or female?
2. Are you introverted or extraverted?
3. On average, how many hours of sleep do you get per night?
4. What is your bedtime: 8pm-10pm, 10pm-12am, 12am-2am, later than 2am?
5. How many countries have you visited?
6. On a scale of 1(very little) to 5 (a lot), how much do you dread this semester?

Data matrix

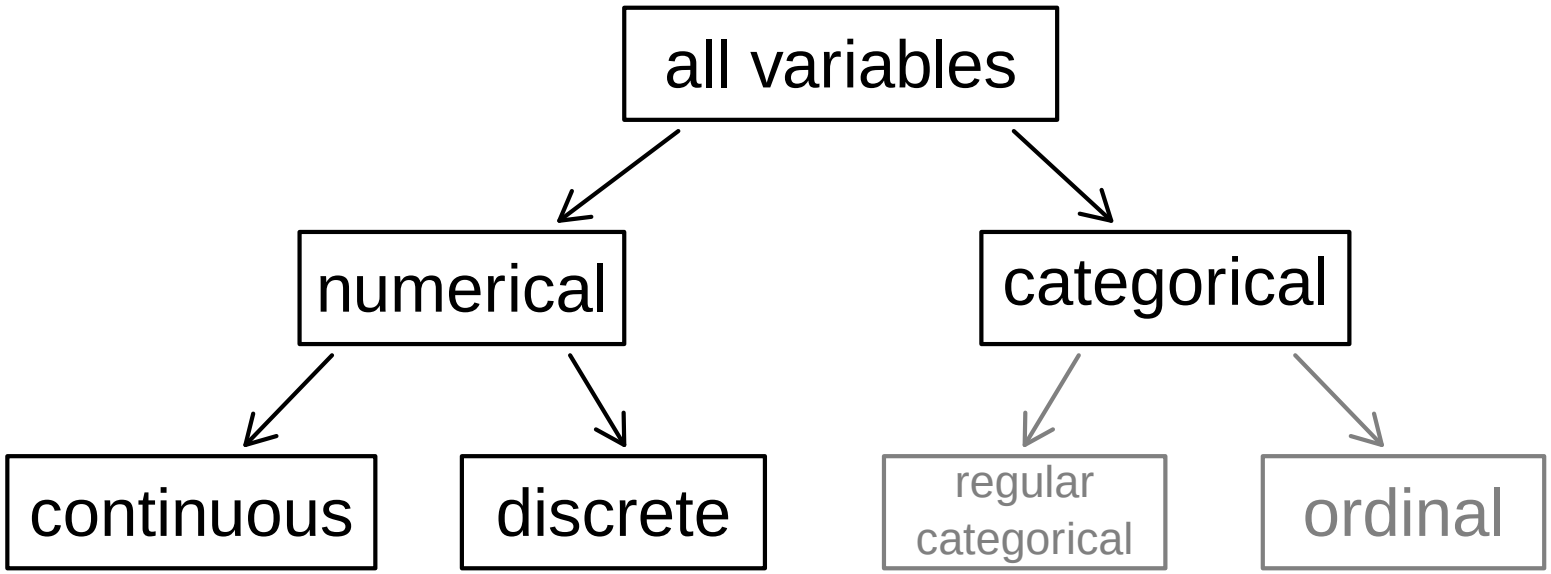
The matrix below shows a sample of responses.

variable ↓

Stu.	gender	intro_extra	···	dread
1	male	extravert	···	3
2	female	extravert	···	2
3	female	introvert	···	4
4	female	extravert	···	2
⋮	⋮	⋮	⋮	⋮
86	male	extravert	···	3

← *observation*

Types of variables



Types of variables (cont.)

	gender	sleep	bedtime	countries	dread
1	male	5	12-2	13	3
2	female	7	10-12	7	2
3	female	5.5	12-2	1	4
4	female	7	12-2		2
5	female	3	12-2	1	3
6	female	3	12-2	9	4

- gender: no inherent order between male and female
- sleep: even though data is reported as whole numbers, sleep is measured on a continuous scale, people just tend to round their responses in surveys
- bedtime: there is an inherent ordering in these time intervals
- countries: data are counted, and can only take on whole numbers
- dread: categories have an inherent ordering

Types of variables (cont.)

	gender	sleep	bedtime	countries	dread
1	male	5	12-2	13	3
2	female	7	10-12	7	2
3	female	5.5	12-2	1	4
4	female	7	12-2		2
5	female	3	12-2	1	3
6	female	3	12-2	9	4

- gender: *categorical* no inherent order between male and female
- sleep: even though data is reported as whole numbers, sleep is measured on a continuous scale, people just tend to round their responses in surveys
- bedtime: there is an inherent ordering in these time intervals
- countries: data are counted, and can only take on whole numbers
- dread: categories have an inherent ordering

Types of variables (cont.)

	gender	sleep	bedtime	countries	dread
1	male	5	12-2	13	3
2	female	7	10-12	7	2
3	female	5.5	12-2	1	4
4	female	7	12-2		2
5	female	3	12-2	1	3
6	female	3	12-2	9	4

- gender: *categorical* no inherent order between male and female
- sleep: even though data is reported as whole numbers, sleep is measured on a continuous scale, people just tend to round their responses in surveys
- bedtime: there is an inherent ordering in these time intervals
- countries: data are counted, and can only take on whole numbers
- dread: categories have an inherent ordering

Types of variables (cont.)

	gender	sleep	bedtime	countries	dread
1	male	5	12-2	13	3
2	female	7	10-12	7	2
3	female	5.5	12-2	1	4
4	female	7	12-2		2
5	female	3	12-2	1	3
6	female	3	12-2	9	4

- gender: *categorical* no inherent order between male and female
- sleep: *numerical, continuous* even though data is reported as whole numbers, sleep is measured on a continuous scale, people just tend to round their responses in surveys
- bedtime: there is an inherent ordering in these time intervals
- countries: data are counted, and can only take on whole numbers
- dread: categories have an inherent ordering

Types of variables (cont.)

	gender	sleep	bedtime	countries	dread
1	male	5	12-2	13	3
2	female	7	10-12	7	2
3	female	5.5	12-2	1	4
4	female	7	12-2		2
5	female	3	12-2	1	3
6	female	3	12-2	9	4

- gender: *categorical* no inherent order between male and female
- sleep: *numerical, continuous* even though data is reported as whole numbers, sleep is measured on a continuous scale, people just tend to round their responses in surveys
- bedtime: there is an inherent ordering in these time intervals
- countries: data are counted, and can only take on whole numbers
- dread: categories have an inherent ordering

Types of variables (cont.)

	gender	sleep	bedtime	countries	dread
1	male	5	12-2	13	3
2	female	7	10-12	7	2
3	female	5.5	12-2	1	4
4	female	7	12-2		2
5	female	3	12-2	1	3
6	female	3	12-2	9	4

- gender: *categorical* no inherent order between male and female
- sleep: *numerical, continuous* even though data is reported as whole numbers, sleep is measured on a continuous scale, people just tend to round their responses in surveys
- bedtime: *categorical, ordinal* there is an inherent ordering in these time intervals
- countries: data are counted, and can only take on whole numbers
- dread: categories have an inherent ordering

Types of variables (cont.)

	gender	sleep	bedtime	countries	dread
1	male	5	12-2	13	3
2	female	7	10-12	7	2
3	female	5.5	12-2	1	4
4	female	7	12-2		2
5	female	3	12-2	1	3
6	female	3	12-2	9	4

- gender: *categorical* no inherent order between male and female
- sleep: *numerical, continuous* even though data is reported as whole numbers, sleep is measured on a continuous scale, people just tend to round their responses in surveys
- bedtime: *categorical, ordinal* there is an inherent ordering in these time intervals
- countries: data are counted, and can only take on whole numbers
- dread: categories have an inherent ordering

Types of variables (cont.)

	gender	sleep	bedtime	countries	dread
1	male	5	12-2	13	3
2	female	7	10-12	7	2
3	female	5.5	12-2	1	4
4	female	7	12-2		2
5	female	3	12-2	1	3
6	female	3	12-2	9	4

- gender: *categorical* no inherent order between male and female
- sleep: *numerical, continuous* even though data is reported as whole numbers, sleep is measured on a continuous scale, people just tend to round their responses in surveys
- bedtime: *categorical, ordinal* there is an inherent ordering in these time intervals
- countries: *numerical, discrete* data are counted, and can only take on whole numbers
- dread: categories have an inherent ordering

Types of variables (cont.)

	gender	sleep	bedtime	countries	dread
1	male	5	12-2	13	3
2	female	7	10-12	7	2
3	female	5.5	12-2	1	4
4	female	7	12-2		2
5	female	3	12-2	1	3
6	female	3	12-2	9	4

- gender: *categorical* no inherent order between male and female
- sleep: *numerical, continuous* even though data is reported as whole numbers, sleep is measured on a continuous scale, people just tend to round their responses in surveys
- bedtime: *categorical, ordinal* there is an inherent ordering in these time intervals
- countries: *numerical, discrete* data are counted, and can only take on whole numbers
- dread: categories have an inherent ordering

Types of variables (cont.)

	gender	sleep	bedtime	countries	dread
1	male	5	12-2	13	3
2	female	7	10-12	7	2
3	female	5.5	12-2	1	4
4	female	7	12-2		2
5	female	3	12-2	1	3
6	female	3	12-2	9	4

- gender: *categorical* no inherent order between male and female
- sleep: *numerical, continuous* even though data is reported as whole numbers, sleep is measured on a continuous scale, people just tend to round their responses in surveys
- bedtime: *categorical, ordinal* there is an inherent ordering in these time intervals
- countries: *numerical, discrete* data are counted, and can only take on whole numbers
- dread: *categorical, ordinal* categories have an inherent ordering

Practice

What type of variable is a telephone area code?

(a) numerical, continuous

(b) numerical, discrete

(c) categorical

(d) categorical, ordinal

Practice

What type of variable is a telephone area code?

(a) numerical, continuous

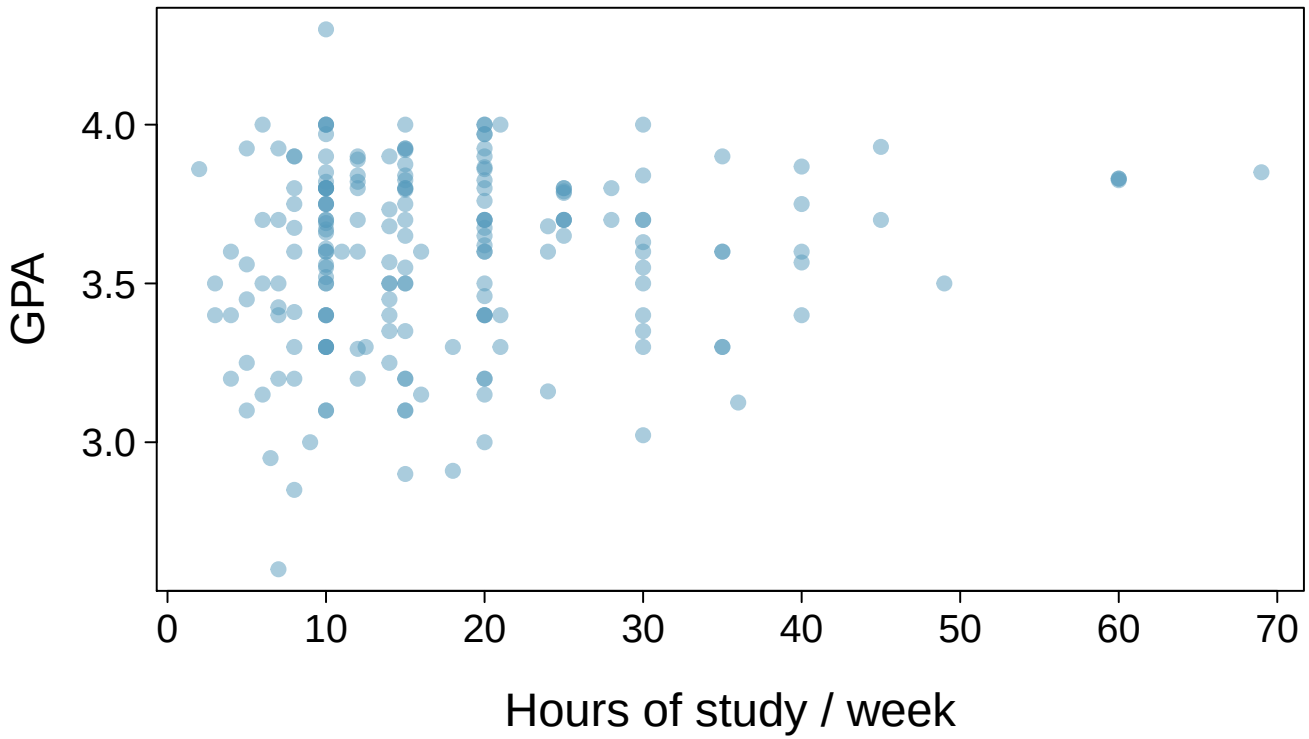
(b) numerical, discrete

(c) *categorical*

(d) categorical, ordinal

Relationships among variables

Does there appear to be a relationship between GPA and number of hours students study per week?

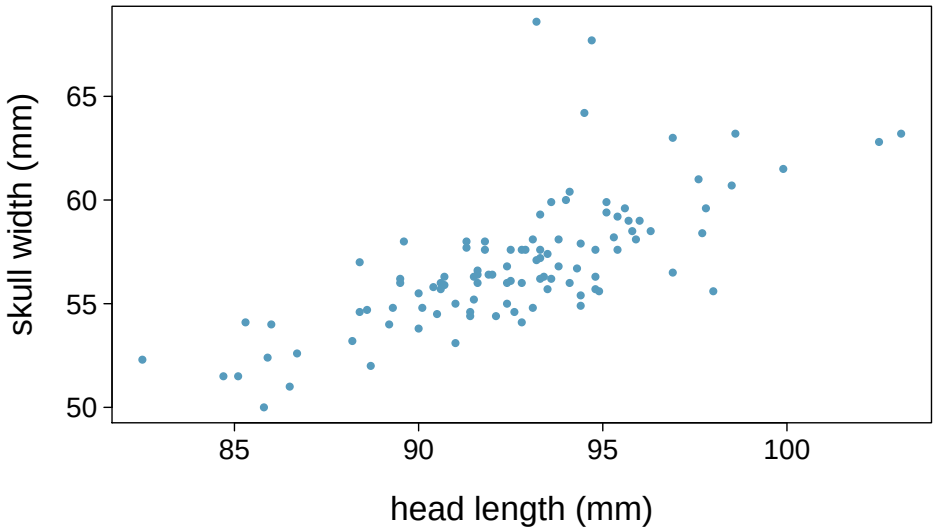


Can you spot anything unusual about any of the data points?

There is one student with GPA > 4.0, this is likely a data error.

Practice

Based on the scatterplot on the right, which of the following statements is correct about the head and skull lengths of possums?



(a) There is no relationship between head length and skull width, i.e. the variables are independent.

(b) Head length and skull width are positively associated.

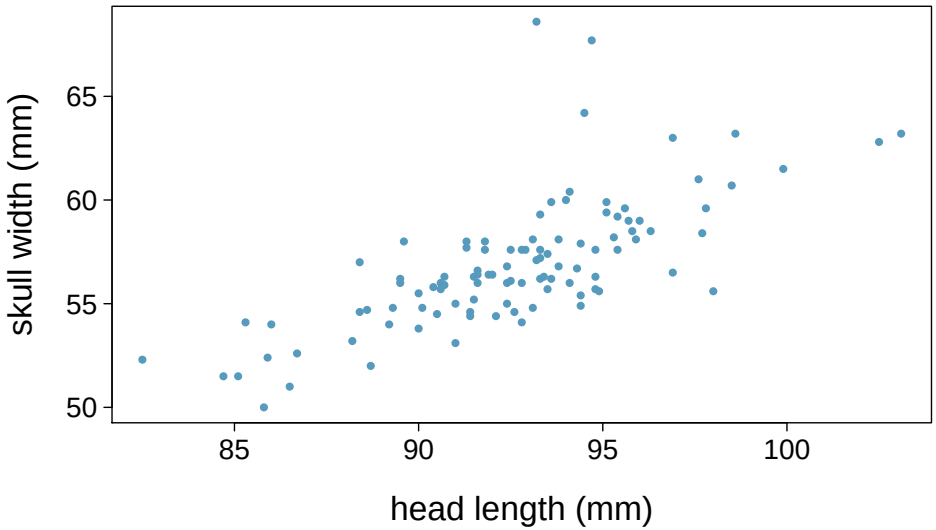
(c) Skull width and head length are negatively associated.

(d) A longer head causes the skull to be wider.

(e) A wider skull causes the head to be longer.

Practice

Based on the scatterplot on the right, which of the following statements is correct about the head and skull lengths of possums?



(a) There is no relationship between head length and skull width, i.e. the variables are independent.

(b) *Head length and skull width are positively associated.*

(c) Skull width and head length are negatively associated.

(d) A longer head causes the skull to be wider.

(e) A wider skull causes the head to be longer.

Associated vs. independent

- When two variables show some connection with one another, they are called *associated* variables.
  - Associated variables can also be called *dependent* variables and vice-versa.
- If two variables are not associated, i.e. there is no evident connection between the two, then they are said to be *independent*.

Populations and samples

PHYS ED | AUGUST 29, 2012, 12:01 AM | 21 Comments

Finding Your Ideal Running Form

By GRETCHEN REYNOLDS

David De Lossy/Getty Images

http://well.blogs.nytimes.com/2012/08/29/finding-your-ideal-running-form

Research question: Can people become better, more efficient runners on their own, merely by running?
Population of interest: *All people*

Sample: Group of adult women who recently joined a running group
Population to which results can be generalized: *Adult women, if the data are randomly sampled*

Anecdotal evidence and early smoking research

- Anti-smoking research started in the 1930s and 1940s when cigarette smoking became increasingly popular. While some smokers seemed to be sensitive to cigarette smoke, others were completely unaffected.
- Anti-smoking research was faced with resistance based on *anecdotal evidence* such as “My uncle smokes three packs a day and he’s in perfectly good health", evidence based on a limited sample size that might not be representative of the population.
- It was concluded that “smoking is a complex human behavior, by its nature difficult to study, confounded by human variability."
- In time researchers were able to examine larger samples of cases (smokers), and trends showing that smoking has negative health impacts became much clearer.

Brandt, *The Cigarette Century* (2009), Basic Books.

Census

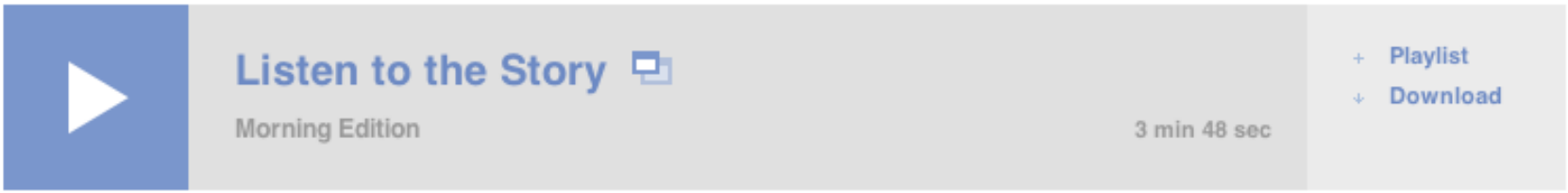
- Wouldn't it be better to just include everyone and "sample" the entire population?
  - This is called a *census*.
- There are problems with taking a census:
  - It can be difficult to complete a census: there always seem to be some individuals who are hard to locate or hard to measure. *And these difficult-to-find people may have certain characteristics that distinguish them from the rest of the population.*
  - Populations rarely stand still. Even if you could take a census, the population changes constantly, so it's never possible to get a perfect measure.
  - Taking a census may be more complex than sampling.

Illegal Immigrants Reluctant To Fill Out Census Form

by PETER O'DOWD

March 31, 2010 4:00 AM

from KJZZ



There is an effort underway to make sure Hispanics are accurately counted in the 2010 Census. Phoenix has some of the country's "hardest-to-count" districts. Some Latinos, especially illegal residents, fear that participating in the count will expose them to immigration raids or government harassment.

http://www.npr.org/templates/story/story.php?storyId=125380052

Exploratory analysis to inference

- Sampling is natural.
- Think about sampling something you are cooking - you taste (examine) a small part of what you're cooking to get an idea about the dish as a whole.
- When you taste a spoonful of soup and decide the spoonful you tasted isn't salty enough, that's *exploratory analysis*.
- If you generalize and conclude that your entire soup needs salt, that's an *inference*.
- For your inference to be valid, the spoonful you tasted (the sample) needs to be *representative* of the entire pot (the population).
  - If your spoonful comes only from the surface and the salt is collected at the bottom of the pot, what you tasted is probably not representative of the whole pot.
  - If you first stir the soup thoroughly before you taste, your spoonful will more likely be representative of the whole pot.

Sampling bias

- *Non-response:* If only a small fraction of the randomly sampled people choose to respond to a survey, the sample may no longer be representative of the population.
- *Voluntary response:* Occurs when the sample consists of people who volunteer to respond because they have strong opinions on the issue. Such a sample will also not be representative of the population.

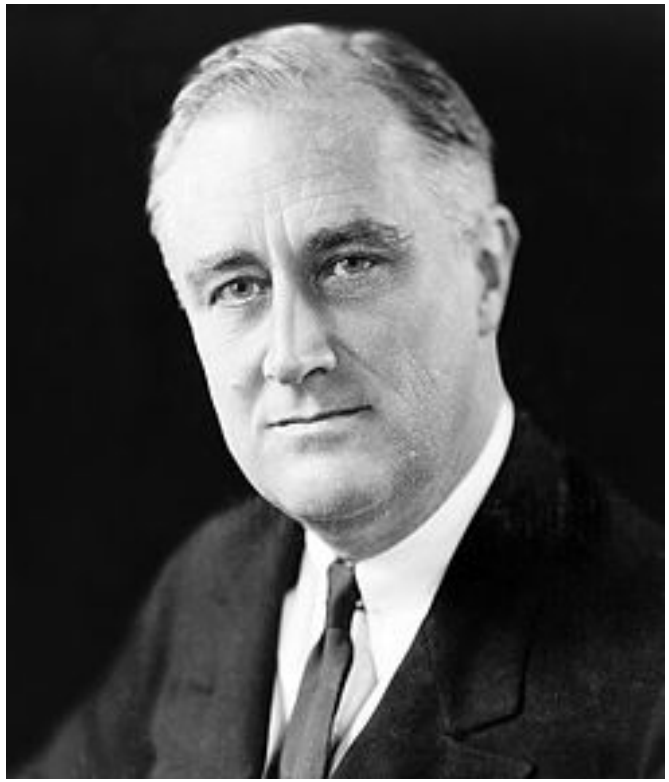
cnn.com, Jan 14, 2012

- *Convenience sample:* Individuals who are easily accessible are more likely to be included in the sample.

Sampling bias example: Landon vs. FDR

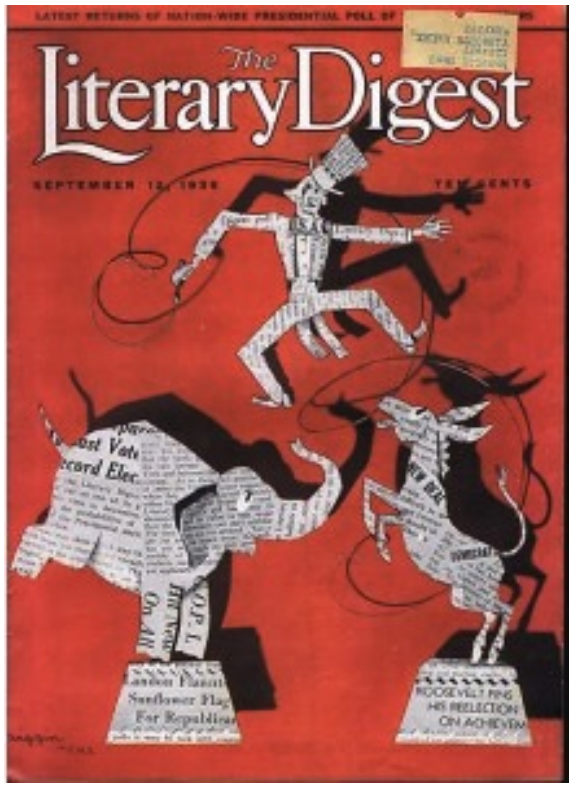
A historical example of a biased sample yielding misleading results:

In 1936, Landon sought the Republican presidential nomination opposing the re-election of FDR.



The Literary Digest Poll

- The Literary Digest polled about 10 million Americans, and got responses from about 2.4 million.
- The poll showed that Landon would likely be the overwhelming winner and FDR would get only 43% of the votes.
- Election result: FDR won, with 62% of the votes.



- The magazine was completely discredited because of the poll, and was soon discontinued.

The Literary Digest Poll – what went wrong?

- The magazine had surveyed
  - its own readers,
  - registered automobile owners, and
  - registered telephone users.
- These groups had incomes well above the national average of the day (remember, this is Great Depression era) which resulted in lists of voters far more likely to support Republicans than a truly *typical* voter of the time, i.e. the sample was not representative of the American population at the time.

Large samples are preferable, but...

- The Literary Digest election poll was based on a sample size of 2.4 million, which is huge, but since the sample was *biased*, the sample did not yield an accurate prediction.
- Back to the soup analogy: If the soup is not well stirred, it doesn't matter how large a spoon you have, it will still not taste right. If the soup is well stirred, a small spoon will suffice to test the soup.

Practice

A school district is considering whether it will no longer allow high school students to park at school after two recent accidents where students were severely injured. As a first step, they survey parents by mail, asking them whether or not the parents would object to this policy change. Of 6,000 surveys that go out, 1,200 are returned. Of these 1,200 surveys that were completed, 960 agreed with the policy change and 240 disagreed. Which of the following statements are true?

I. Some of the mailings may have never reached the parents.

II. The school district has strong support from parents to move forward with the policy approval.

III. It is possible that majority of the parents of high school students disagree with the policy change.

IV. The survey results are unlikely to be biased because all parents were mailed a survey.

(a) Only I (b) I and II (c) I and III (d) III and IV (e) Only IV

Practice

A school district is considering whether it will no longer allow high school students to park at school after two recent accidents where students were severely injured. As a first step, they survey parents by mail, asking them whether or not the parents would object to this policy change. Of 6,000 surveys that go out, 1,200 are returned. Of these 1,200 surveys that were completed, 960 agreed with the policy change and 240 disagreed. Which of the following statements are true?

I. Some of the mailings may have never reached the parents.

II. The school district has strong support from parents to move forward with the policy approval.

III. It is possible that majority of the parents of high school students disagree with the policy change.

IV. The survey results are unlikely to be biased because all parents were mailed a survey.

(a) Only I (b) I and II (c) *I and III* (d) III and IV (e) Only IV

Explanatory and response variables

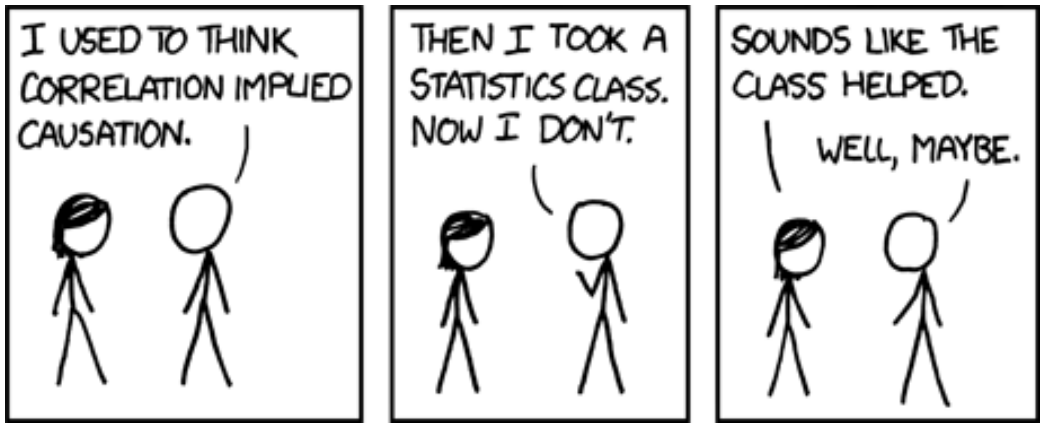
- To identify the explanatory variable in a pair of variables, identify which of the two is suspected of affecting the other:

$$\text{explanatory variable} \xrightarrow{\text{might affect}} \text{response variable}$$

- Labeling variables as explanatory and response does not guarantee the relationship between the two is actually causal, even if there is an association identified between the two variables. We use these labels only to keep track of which variable we suspect affects the other.

Observational studies and experiments

- *Observational study:* Researchers collect data in a way that does not directly interfere with how the data arise, i.e. they merely "observe", and can only establish an association between the explanatory and response variables.
- *Experiment:* Researchers randomly assign subjects to various treatments in order to establish causal connections between the explanatory and response variables.
- If you're going to walk away with one thing from this class, let it be **"correlation does not imply causation"**.



http://xkcd.com/552/

New study sponsored by General Mills says that eating breakfast makes girls thinner

Study: Breakfast Helps Girls Stay Slim
I love these studies....and finding out who sponsored them!

By ALEX DOMINGUEZ, Associated Press

Girls who regularly ate breakfast, particularly one that includes cereal, were slimmer than those who skipped the morning meal, according to a study that tracked nearly 2,400 girls for 10 years.

Girls who ate breakfast of any type had a lower average body mass index, a common obesity gauge, than those who said they didn't. The index was even lower for girls who said they ate cereal for breakfast, according to findings of the study conducted by the Maryland Medical Research Institute. The study received funding from the National Institutes of Health and cereal-maker General Mills.

"Not eating breakfast is the worst thing you can do, that's really the take-home message for teenage girls," said study author Bruce Barton, the Maryland institute's president and CEO.

The fiber in cereal and healthier foods that normally accompany cereal, such as milk and orange juice, may account for the lower body mass index among cereal eaters, Barton said.

The results were gleaned from a larger NIH survey of 2,379 girls in California, Ohio and Maryland who were tracked between ages 9 and 19. Results of the study appear in the September issue of the Journal of the American Dietetic Association.

Nearly one in three adolescent girls in the United States is overweight, according to the association. The problem is particularly troubling because research shows becoming overweight as a child can lead to a lifetime struggle with obesity.

As part of the survey, the girls were asked once a year what they had eaten during the previous three days. The data were adjusted to compensate for factors such as differences in physical activity among the girls and normal increases in body fat during adolescence.

What type of study is this, observational study or an experiment? *"Girls who regularly ate breakfast, particularly one that includes cereal, were slimmer than those who skipped the morning meal, according to a study that tracked nearly 2,400 girls for 10 years. [...] As part of the survey, the girls were asked once a year what they had eaten during the previous three days."*

What is the conclusion of the study?

Who sponsored the study?

What type of study is this, observational study or an experiment? *"Girls who regularly ate breakfast, particularly one that includes cereal, were slimmer than those who skipped the morning meal, according to a study that tracked nearly 2,400 girls for 10 years. [...] As part of the survey, the girls were asked once a year what they had eaten during the previous three days."*

This is an observational study since the researchers merely observed the behavior of the girls (subjects) as opposed to imposing treatments on them.

What is the conclusion of the study?

Who sponsored the study?

What type of study is this, observational study or an experiment? *"Girls who regularly ate breakfast, particularly one that includes cereal, were slimmer than those who skipped the morning meal, according to a study that tracked nearly 2,400 girls for 10 years. [...] As part of the survey, the girls were asked once a year what they had eaten during the previous three days."*

This is an observational study since the researchers merely observed the behavior of the girls (subjects) as opposed to imposing treatments on them.

What is the conclusion of the study?

There is an association between girls eating breakfast and being slimmer.

Who sponsored the study?

What type of study is this, observational study or an experiment? *"Girls who regularly ate breakfast, particularly one that includes cereal, were slimmer than those who skipped the morning meal, according to a study that tracked nearly 2,400 girls for 10 years. [...] As part of the survey, the girls were asked once a year what they had eaten during the previous three days."*

This is an observational study since the researchers merely observed the behavior of the girls (subjects) as opposed to imposing treatments on them.

What is the conclusion of the study?

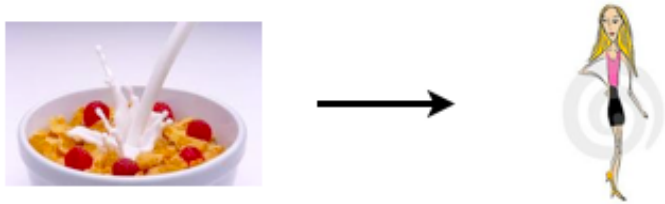
There is an association between girls eating breakfast and being slimmer.

Who sponsored the study?

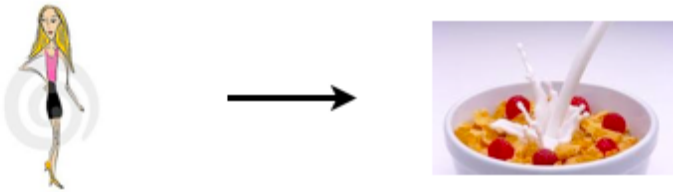
General Mills.

3 possible explanations

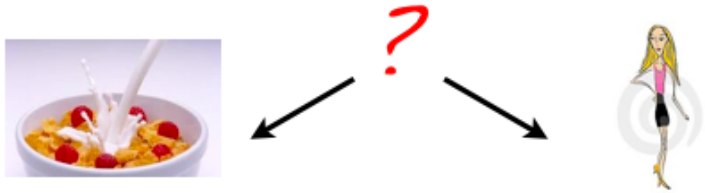
1. Eating breakfast causes girls to be thinner.



2. Being thin causes girls to eat breakfast.



3. A third variable is responsible for both. What could it be?
   An extraneous variable that affects both the explanatory and the response variable and that make it seem like there is a relationship between the two are called *confounding* variables.



Images from: *http://www.appforhealth.com/wp-content/uploads/2011/08/ipn-cerealfrijo-300x135.jpg*, *http://www.dreamstime.com/stock-photography-too-thin-woman-anorexia-model-image2814892*.

Prospective vs. retrospective studies

- A *prospective* study identifies individuals and collects information as events unfold.
  - Example: The Nurses Health Study has been recruiting registered nurses and then collecting data from them using questionnaires since 1976.
- *Retrospective studies* collect data after events have taken place.
  - Example: Researchers reviewing past events in medical records.

Principles of experimental design

1. *Control:* Compare treatment of interest to a control group.
2. *Randomize:* Randomly assign subjects to treatments, and randomly sample from the population whenever possible.
3. *Replicate:* Within a study, replicate by collecting a sufficiently large sample. Or replicate the entire study.
4. *Block:* If there are variables that are known or suspected to affect the response variable, first group subjects into *blocks* based on these variables, and then randomize cases within each block to treatment groups.

More on blocking

- We would like to design an experiment to investigate if energy gels makes you run faster:
  - Treatment: energy gel
  - Control: no energy gel
- It is suspected that energy gels might affect pro and amateur athletes differently, therefore we block for pro status:
  - Divide the sample to pro and amateur
  - Randomly assign pro athletes to treatment and control groups
  - Randomly assign amateur athletes to treatment and control groups
  - Pro/amateur status is equally represented in the resulting treatment and control groups

Why is this important? Can you think of other variables to block for?

Practice

A study is designed to test the effect of light level and noise level on exam performance of students. The researcher also believes that light and noise levels might have different effects on males and females, so wants to make sure both genders are equally represented in each group. Which of the below is correct?

(a) There are 3 explanatory variables (light, noise, gender) and 1 response variable (exam performance)

(b) There are 2 explanatory variables (light and noise), 1 blocking variable (gender), and 1 response variable (exam performance)

(c) There is 1 explanatory variable (gender) and 3 response variables (light, noise, exam performance)

(d) There are 2 blocking variables (light and noise), 1 explanatory variable (gender), and 1 response variable (exam performance)

Practice

A study is designed to test the effect of light level and noise level on exam performance of students. The researcher also believes that light and noise levels might have different effects on males and females, so wants to make sure both genders are equally represented in each group. Which of the below is correct?

(a) There are 3 explanatory variables (light, noise, gender) and 1 response variable (exam performance)

(b) *There are 2 explanatory variables (light and noise), 1 blocking variable (gender), and 1 response variable (exam performance)*

(c) There is 1 explanatory variable (gender) and 3 response variables (light, noise, exam performance)

(d) There are 2 blocking variables (light and noise), 1 explanatory variable (gender), and 1 response variable (exam performance)

Difference between blocking and explanatory variables

- Factors are conditions we can impose on the experimental units.
- Blocking variables are characteristics that the experimental units come with, that we would like to control for.

More experimental design terminology...

- *Placebo:* fake treatment, often used as the control group for medical studies
- *Placebo effect:* experimental units showing improvement simply because they believe they are receiving a special treatment
- *Blinding:* when experimental units do not know whether they are in the control or treatment group
- *Double-blind:* when both the experimental units and the researchers who interact with the patients do not know who is in the control and who is in the treatment group

Practice

What is the main difference between observational studies and experiments?

(a) Experiments take place in a lab while observational studies do not need to.

(b) In an observational study we only look at what happened in the past.

(c) Most experiments use random assignment while observational studies do not.

(d) Observational studies are completely useless since no causal inference can be made based on their findings.

Practice

What is the main difference between observational studies and experiments?

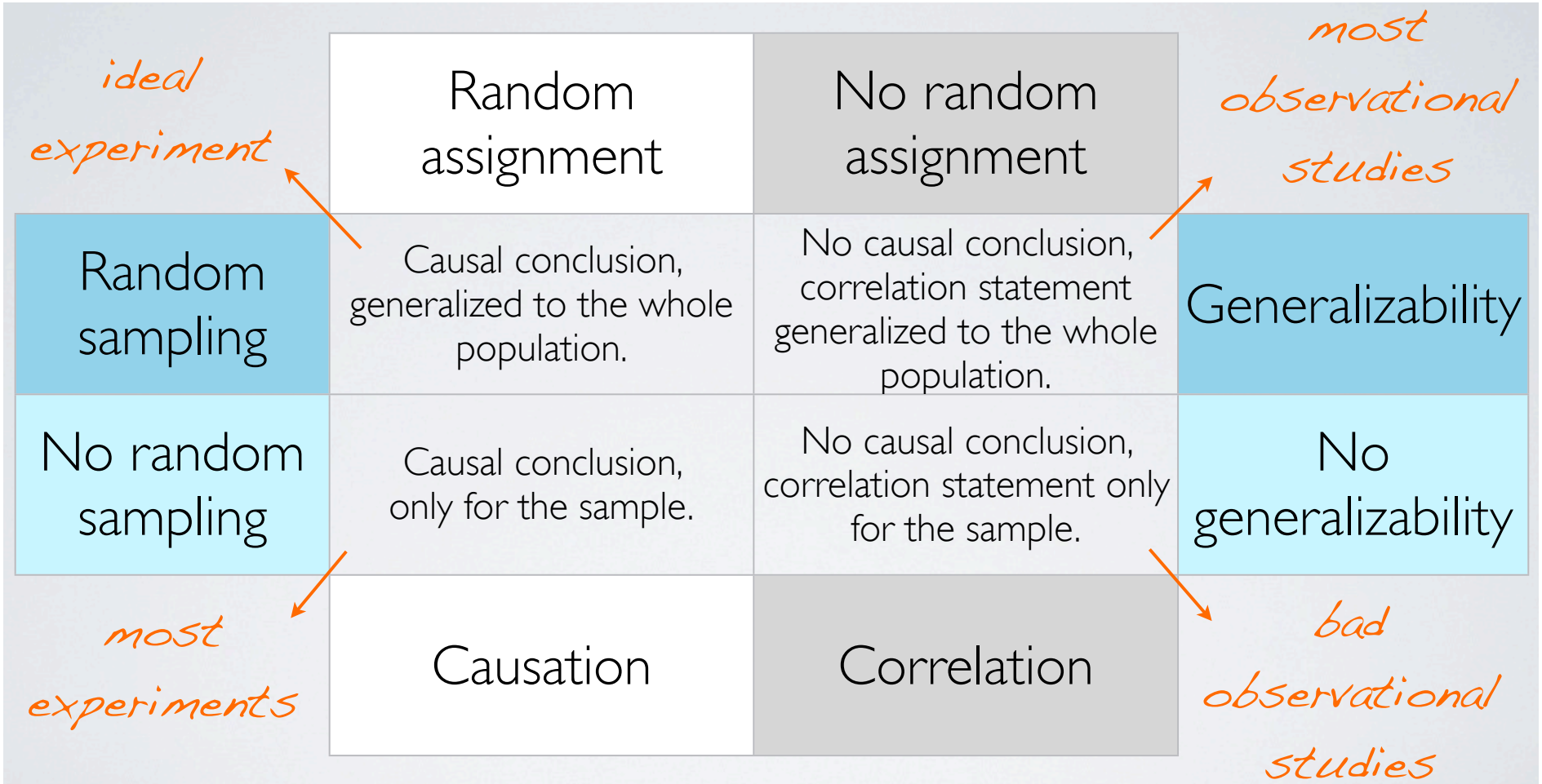
(a) Experiments take place in a lab while observational studies do not need to.

(b) In an observational study we only look at what happened in the past.

(c) *Most experiments use random assignment while observational studies do not.*

(d) Observational studies are completely useless since no causal inference can be made based on their findings.

Random assignment vs. random sampling



Historical controls

- **Historical controls** do not give a randomized experiment, which is one reason their use is problematic. The FDA is very reluctant to approve drugs in which all patients in the trial receive the drug, while the control group are patients who were treated before the drug was invented. One concern is that the standard of basic care constantly improves, so the drug may appear effective when, in fact, the only difference is that current patients get, say, better nursing care.

Confounding Factors

- In an observational study, the researcher does not get to determine who receives the treatment. For example, people who smoke get lung cancer at a higher rate than those who do not smoke. Does smoking cause lung cancer?

  The tobacco lobby used to say no, arguing that: there might be a gene that predisposes people to both enjoy smoking and get cancer; people who like to smoke may tend to follow unhealthy lifestyles (e.g., alcohol use), and that may be the real cause of lung cancer; no randomized, controlled, double-blind experiment (on humans) has shown causation.

Confounding Factors (Cont'd)

- Obviously, it would be ethically problematic to do a randomized controlled experiment (one would have to assign 14 year-olds at random to smoke heavily for the rest of their lives). And it would be hard to make this double-blind—people know if they smoke.

  But animal studies strongly indicate that smoking causes lung cancer in mammals and birds.

- The other two arguments from the tobacco lobby carry more weight. The differences between lung cancer rates in the smokers and non-smokers may be due to smoking, or they may be due to a confounding factor or variable.

  In this case, tobacco lobbies suggested two possible confounding factors: genes and lifestyle.

Confounding Factors (Cont'd)

- One way to try to handle confounding is to make subgroup comparisons that control for possible confounding effects. For example, one could compare the lung cancer rates for smokers who use matches against smokers who use lighters.
- Do seatbelts save lives?

  Seatbelt studies are usually observational (why?). One compares the fatality rates in accidents in which seatbelts were worn to the fatality rate in accidents without seatbelts.

  But one has to worry about confounding factors. For example,
  - People who don't wear seatbelts may drive more recklessly.
  - People who don't wear seatbelts may prefer cars that are not designed with safety in mind.

Confounding Factors (Cont'd)

- Some researchers try to control for this by comparing the fatality rates among seatbelt wearers and non-wearers in similar cars, or cars that are thought to have been traveling at the same speed. But this is awkward to do and invites criticism.
- In order to control for a confounding factor, one has to guess what it is. But that can be hard and you are never sure that you have thought of everything.
- In contrast, with a **randomized design**, the random assignment of people to the treatment and control groups ensures that there is almost no chance of a systematic difference between the groups.

Example: UC Berkeley gender bias

- One of the best-known examples of **Simpson's paradox** is a study of gender bias among graduate school admissions to University of California, Berkeley.
- The admission figures for the fall of 1973 showed that men applying were more likely than women to be admitted, and the difference was so large that it was unlikely to be due to chance.

	Applicants	Admitted
Men	8442	44%
Women	4321	35%

- But when the Dean asked each department to report their admission rates separately, it turned out that most department accepted a larger proportion of women than men.

Example: UC Berkeley gender bias (Cont'd)

- When examining the individual departments, it appeared that six out of 85 departments were significantly biased against men, whereas only four were significantly biased against women. In fact, the pooled and corrected data showed a "small but statistically significant bias in favor of women."
- The data from the six largest departments are listed below.

Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	**82%**
B	560	63%	25	**68%**
C	325	**37%**	593	34%
D	417	33%	375	**35%**
E	191	**28%**	393	24%
F	373	6%	341	**7%**

Example: UC Berkeley gender bias (Cont'd)

This apparent reversal of a pattern is sometimes called **Simpson's Paradox**. A trend that appears in different groups of data disappears when these groups are combined, and the reverse trend appears for the aggregate data.

Figure: Simpson's Paradox: Is your data tricking you?

Example: UC Berkeley gender bias (Cont'd)

- The Dean asked Professor Betty Scott to study the problem. She showed that women tended to apply to the majors that were most selective, whereas the men applied to majors that were less selective. So overall, the women had higher rejection rates.
- Women tended to apply to competitive departments with low rates of admission even among qualified applicants (such as in the English Department), whereas men tended to apply to less-competitive departments with high rates of admission among the qualified applicants (such as in engineering and chemistry).

Weighted Averages

To put such comparisons on a fair footing, she calculated the weighted average admission rates for women and men, where the weights are determined by the proportion of people applying to each of the different majors. This controls for the confounding variable.

To see how the weighted average works, we focus on just two majors. Assume major A accepts 80% of all applicants, but Major B accepts just 10%. Suppose 100 men and 200 women apply. Consider two scenarios:

Weighted Averages (Cont'd)

Scenario 1: Half the men and half the women apply to A, the rest apply to B.

Scenario 2: 90 men apply to A, 10 to B; but 180 women apply to B, 20 to A.

In the first case, major is not a confounding variable. Men and women show the same major preferences. (Note: They do not have to apply in 50-50 ratios—it would still not be a confounder if both genders applied in 25-75 ratios, for example.)

In the second case, major is a confounder. Men prefer A, but women prefer B.

Weighted Averages (Cont'd)

In Scenario 1, the raw number of men who are accepted is

$$.8 * 50 + .1 * 50 = 45$$

and for women the percentage is the same: (.8 * 100 + .1 * 100)/200 is 45%.

In Scenario 2, the raw number of men who are accepted is

$$.8 * 90 + .1 * 10 = 73$$

or **73%**. And the raw number of women accepted is

$$.1 * 180 + .8 * 20 = 34$$

so their acceptance rate is 34/200 or **17%**. This looks like gender bias, but actually it is not—the admission policy is completely gender blind.

Weighted Averages (Cont'd)

To make a fair comparison, weight the acceptance rates for men in each major by the fraction of people applying to that major:

$$\frac{90+20}{300} * \frac{72}{90} + \frac{10+180}{300} * \frac{1}{10} = .357$$

and the weighted average proportion of women accepted is

$$\frac{90+20}{300} * \frac{16}{20} + \frac{10+180}{300} * \frac{18}{180} = .357$$

The weighted average shows that the acceptance rates for men and women, controlling for major, are equal.

Weighted Averages (Cont'd)

The general formula for finding the weighted average correction for the acceptance rate of men is:

$$\text{wtd avg} = \sum_i \quad (\text{prop. of people applying to major } i) * (\text{acceptance rate for men at major } i)$$

Recap

- Data Basics
- Observational Studies
- Designed Experiments
- Simpson's Paradox

Suggested reading:

- OpenIntro3: Sec. 1.1, 1.2, 1.3, 1.4, 1.5