OpenIntro

Chapter 1: Introduction to data

OpenIntro Statistics, 3rd Edition

Introduction

Goals of The Course

After theory (what you learnin substantive courses) and after design (what you learn in 700), comes analysis (what you learn here and in 702) - telling your story with data.

- Does what you predict actually happen?
- Do your data/does your design meet the assumptions of the proceure you're using?
- How big are the effects?
- How do you convey these results to your readers?

Case study

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.

Data basics

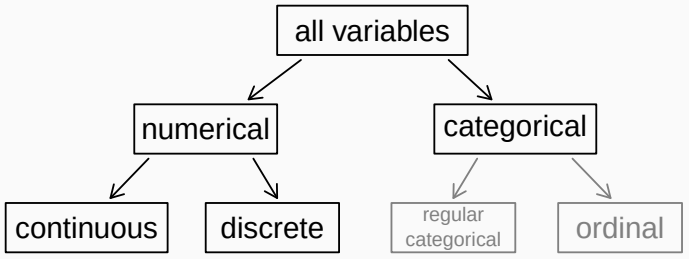
Data matrix

Data collected on students in a statistics class on a variety of variables:

variable ↓

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

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`:
- `sleep`:
- `bedtime`:
- `countries`:
- `dread`:

Practice

What type of variable is a telephone area code?

(a) numerical, continuous

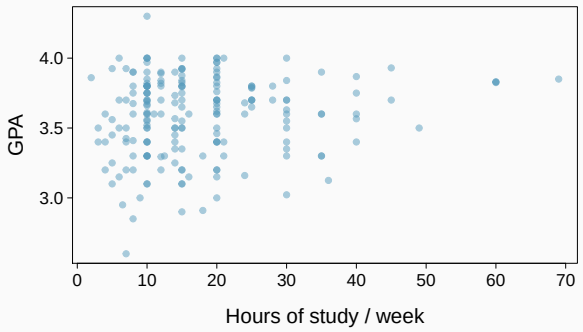
(b) numerical, discrete

(c) categorical

(d) categorical, ordinal

Relationships among variables

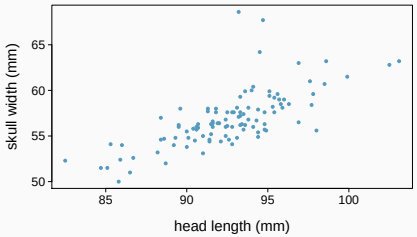
Does there appear to be a relationship between GPA and number of hours students study per week?



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

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*.

Overview of data collection principles

Populations and samples

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:

Sample: Group of adult women who recently joined a running group

Population to which results can be generalized:

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.

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

Listen to the Story

Morning Edition

3 min 48 sec

Playlist

Download

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.

Quick vote

Do you get paid sick days at your job?

Yes No

What job?

VOTE or view results

Quick vote

Do you get paid sick days at your job?

Read Related Articles

Yes 63% 20056

No 21% 6816

What job? 15% 4885

Total votes: 31757

This is not a scientific poll

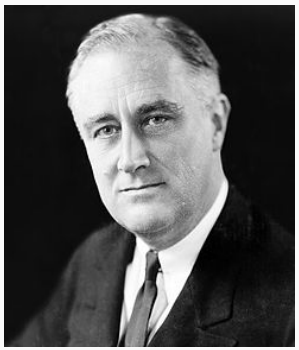
cnn.com, Jan 14, 2012

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

Sampling bias example: Landon vs. FDR

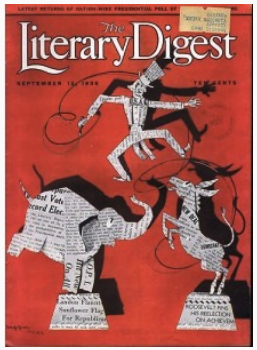
A historical example of a biased sample yielding misleading results:

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

Explanatory and response variables

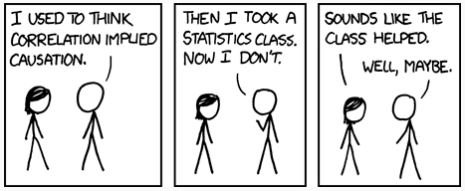
- To identify the explanatory variable in a pair of variables, identify which of the two is suspected of affecting the other:

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

- 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/

Observational studies and sampling strategies

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.

http://www.peertrainer.com/LoungeCommunityThread.aspx?ForumID=1&ThreadID=3118

What type of study is this, observational study or an experiment?

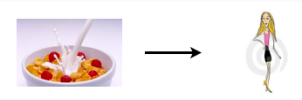
"Girls who regularly ate breakfast, particularly one that includes cereal, were slimmer than those who skipped the morning meal, according to a study that tracked nearly 2,400 girls for 10 years. [...] As part of the survey, the girls were asked once a year what they had eaten during the previous three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

What is the conclusion of the study?

Who sponsored the study?

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

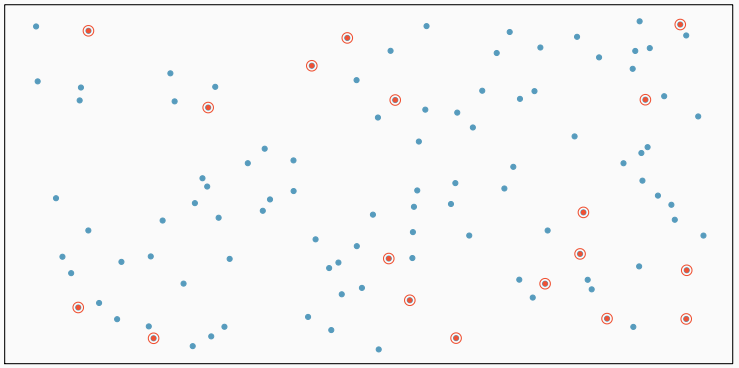
- A *prospective* study identifies individuals and collects information as events unfold.
  - Example: The Nurses Health Study has been recruiting registered nurses and then collecting data from them using questionnaires since 1976.
- *Retrospective studies* collect data after events have taken place.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

Obtaining good samples

- Almost all statistical methods are based on the notion of implied randomness.
- If observational data are not collected in a random framework from a population, these statistical methods – the estimates and errors associated with the estimates – are not reliable.
- Most commonly used random sampling techniques are *simple*, *stratified*, and *cluster* sampling.

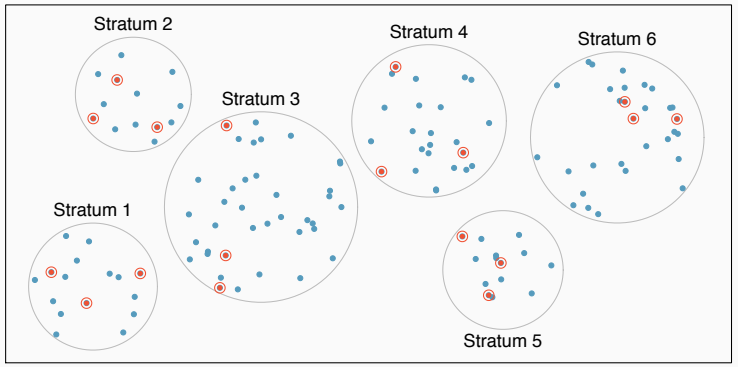
Simple random sample

Randomly select cases from the population, where there is no implied connection between the points that are selected.



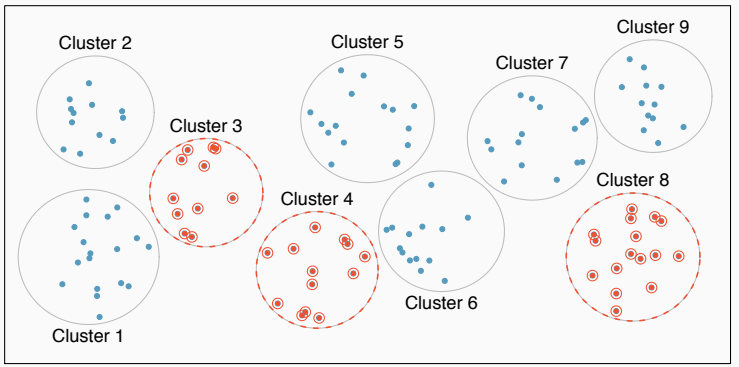
Stratified sample

Strata are made up of similar observations. We take a simple random sample from each stratum.



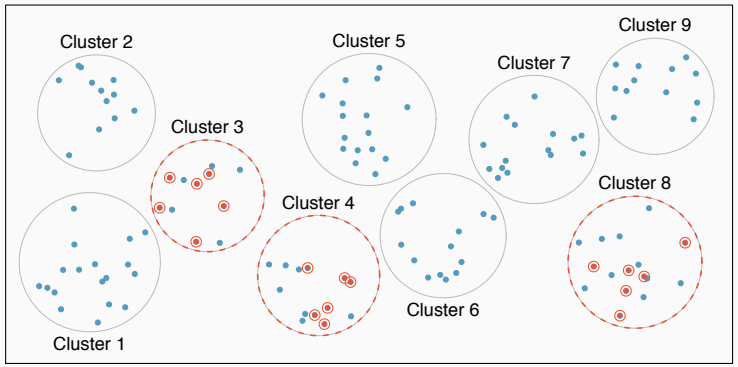
Cluster sample

Clusters are usually not made up of homogeneous observations. We take a simple random sample of clusters, and then sample all observations in that cluster. Usually preferred for economical reasons.



Multistage sample

Clusters are usually not made up of homogeneous observations. We take a simple random sample of clusters, and then take a simple random sample of observations from the sampled clusters.



Practice

A city council has requested a household survey be conducted in a suburban area of their city. The area is broken into many distinct and unique neighborhoods, some including large homes, some with only apartments. Which approach would likely be the least effective?

(a) Simple random sampling

(b) Cluster sampling

(c) Stratified sampling

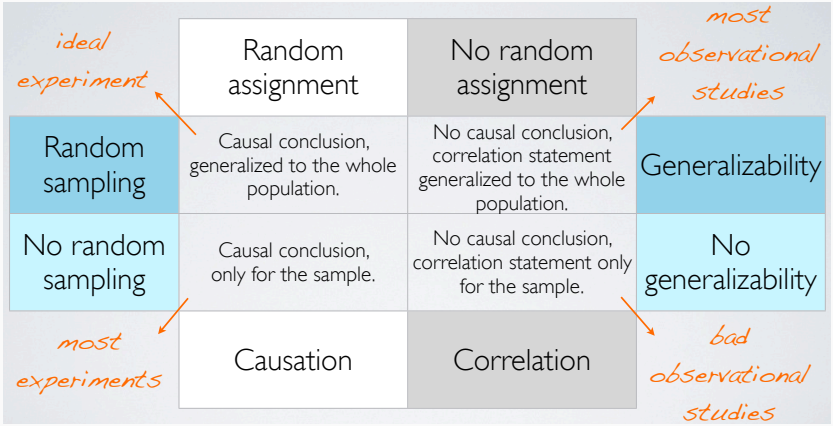
(d) Blocked sampling

Experiments

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.

Random assignment vs. random sampling



Introduction to R

Software Introduction: R

- **R** can be downloaded from `http://cran.r-project.org`
  - For Windows, click on `Download R for Windows`, then click on `base` and finally on `Download R 3.3.1 for Windows`.
    - I would recommend MDI mode when you have the option, but it's up to you.
  - For Mac, click on `Download R for MacOS X`, then on `R-3.3.1-os.pkg` (where os = (snowleopard, mavericks)).
    - You should also go on that same page to the `tools` link and download and install `gfortran-4.2.3.dmg` and `tcltk-8.5.5-x11.dmg`.
- Then, double-clicking the R icon will launch R.
- We will interact with R mostly through RStudio, though.

RStudio

Rstudio is an IDE (Integrated Development Environment) for R. It allows you to:

- write and save code
- view output
- manage your workspace
- even write papers if you want

You can download RStudio from `http://www.rstudio.com/products/rstudio/download/`.

Entering Data in R

You can enter data into R in a bunch of different ways, but I'll talk about a couple. First, you can enter it directly.

```
x <- c(2,3,7,10,11)
mean(x)

## [1] 6.6

median(x)

## [1] 7
```

- You could also enter it into a spreadsheet in excel, save the file as a `.csv`, and then read into R using the `read.csv`.
- You can also read Stata datasets into R and write Stata datasets out of R, but we will save these for later.

Math to Variables: R

In R you can directly to any mathematical operation to any object that is numerical (e.g., a variable). Using `x` above, we could square it (the ^ means "to the power of"). You can add, subtract, multiply and divide with +, -, * and /, respectively. You can also do multiple operations at once.

```
x + 3

## [1]  5  6 10 13 14

x/2

## [1] 1.0 1.5 3.5 5.0 5.5

x*1.5

## [1]  3.0  4.5 10.5 15.0 16.5

(x+5)^3

## [1]  343  512 1728 3375 4096
```


Fixing Mistakes

Let's say that we had an entry that was wrong and we wanted to fix it, let's say the 4^{th} value of `x` was supposed to be 9 instead of 10.

```
x[4] <- 9
summary(x)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     2.0     3.0     7.0     6.4     9.0    11.0
```


Calculating Number of Times Something Happens: R

Calculating the number of times a condition is met in R is easy. For example, if we wanted to know how often x is bigger than 4, we could do.

```
sum(x > 4)

## [1] 3
```

Calculating Percentages: R

Calculating a proportion is just taking the number of times something happens over the total number of times it could have happened. For example, if we wanted to know the proportion of times x is bigger than 4, we need to divide the number of times x is bigger than 4 by the number of values in x. x is bigger than four, 3 times and not bigger than four 2 times.

```
sum(x > 4)/length(x)

## [1] 0.6

mean(x > 4)

## [1] 0.6
```