

Chapter 1: Introduction to data

OpenIntro Statistics, 3rd Edition

Slides developed by Mine Çetinkaya-Rundel of OpenIntro.

The slides may be copied, edited, and/or shared via the CC BY-SA license.

Some images may be included under fair use guidelines (educational purposes).

Case study

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.

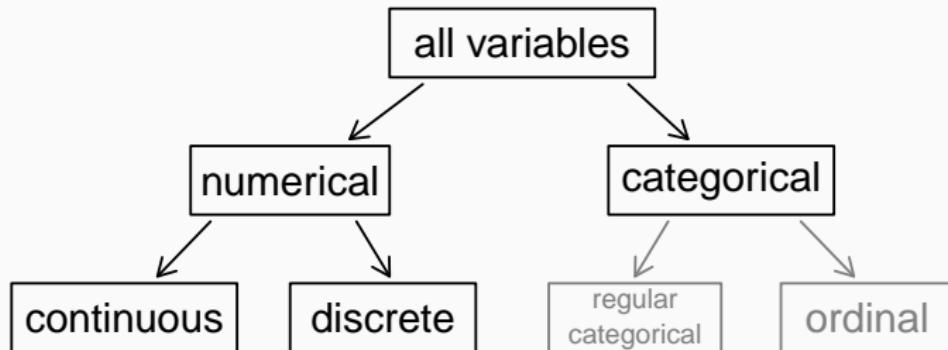
Data basics

Data matrix

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

<i>variable</i>					
Stu.	gender	intro_extra	...	dread	
1	male	extravert	...	3	
2	female	extravert	...	2	
3	female	introvert	...	4	←
4	female	extravert	...	2	<i>observation</i>
:	:	:	:	:	
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: *categorical*
- sleep: *numerical, continuous*
- bedtime: *categorical, ordinal*
- countries: *numerical, discrete*
- dread: *categorical, ordinal - could also be used as numerical*

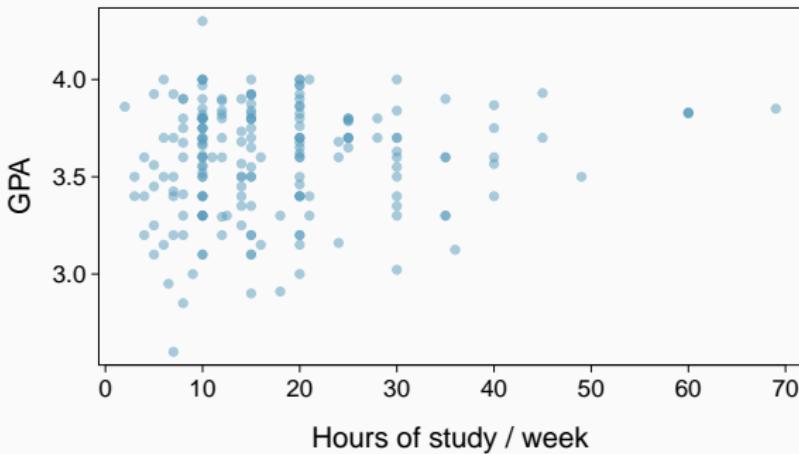
Practice

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?

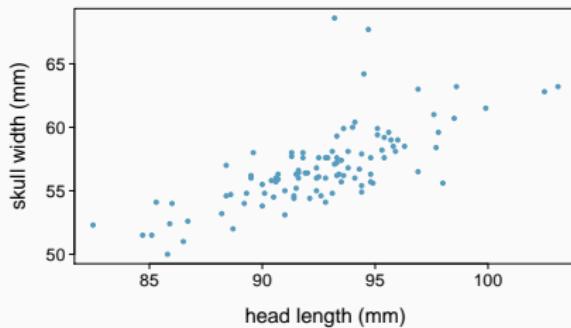


Can you spot anything unusual about any of the data points?

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.

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

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

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

Research question: Can people become better, more efficient runners on their own, merely by running?

Population of interest: All people

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

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:

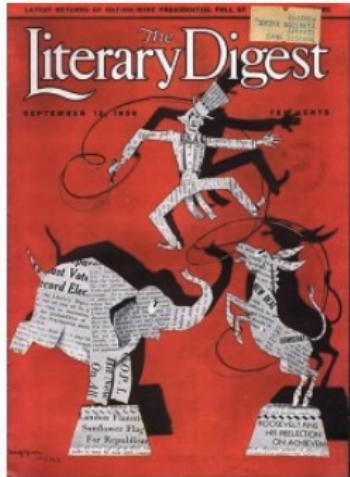


In 1936, Landon sought the Republican presidential nomination opposing the re-election of FDR.



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.

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:

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

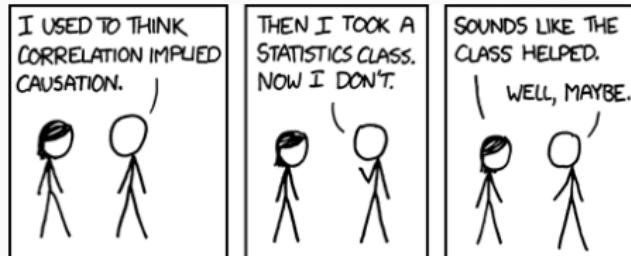
- *Observational study*: Researchers collect data in a way that does not directly interfere with how the data arise, i.e. they merely “observe”, and can only establish an association between the explanatory and response variables.

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.

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”.



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.

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?

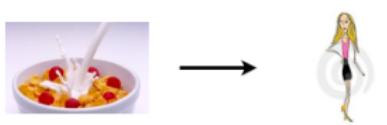
There is an **association** between girls eating breakfast and being slimmer.

Who sponsored the study?

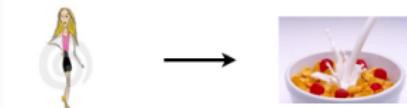
General Mills.

3 possible explanations

1. Eating breakfast causes girls to be thinner.

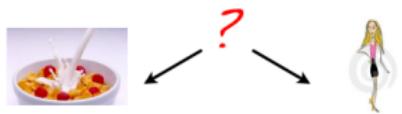


2. Being thin causes girls to eat breakfast.



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/iphn-cerealfrijoles-300x135.jpg>,

<http://www.dreamstime.com/stock-photography-too-thin-woman-anorexia-model-image2814892>.

Prospective vs. retrospective studies

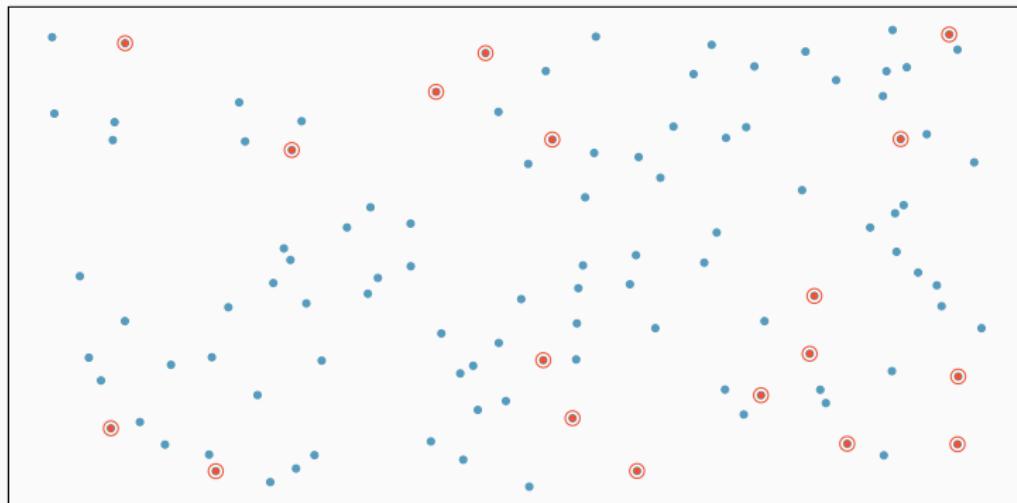
- A *prospective* study identifies individuals and collects information as events unfold.
 - Example: The Nurses Health Study has been recruiting registered nurses and then collecting data from them using questionnaires since 1976.
- *Retrospective studies* collect data after events have taken place.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

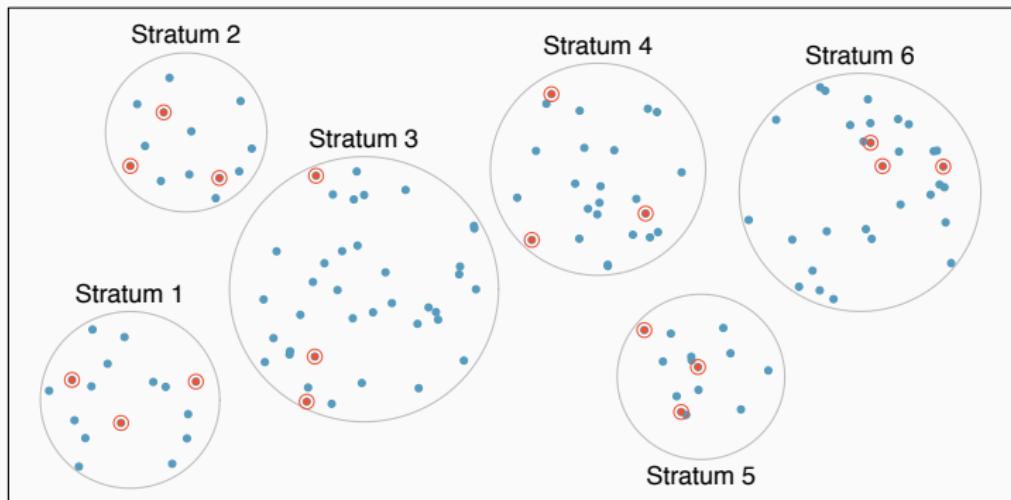
Simple random sample

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



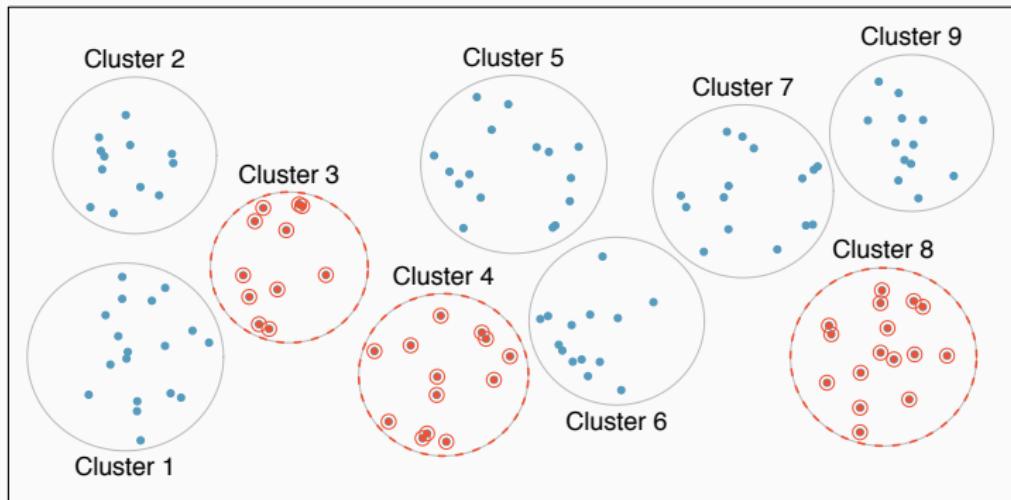
Stratified sample

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



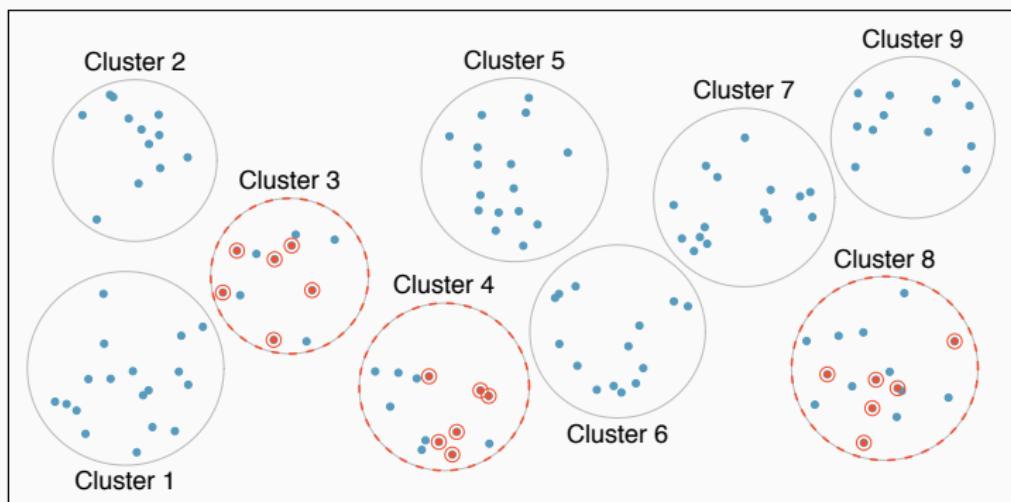
Cluster sample

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.



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.

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

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.
- Blocking is like stratifying, except used in experimental settings when randomly assigning, as opposed to when sampling.

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

Random assignment vs. random sampling

ideal experiment	Random assignment	No random assignment	most observational studies
Random sampling	Causal conclusion, generalized to the whole population.	No causal conclusion, correlation statement generalized to the whole population.	Generalizability
No random sampling	Causal conclusion, only for the sample.	No causal conclusion, correlation statement only for the sample.	No generalizability
most experiments	Causation	Correlation	bad observational studies