Chapter 1: Introduction to data

STA 101, Summer I 2021, Duke University

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. [1]

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

[1] 7 patients dropped out of the study: 3 from the treatment and 4 from the control group.

Results

The table below shows the distribution of patients with good outcomes at 6-month follow-up.

		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:

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

Results

The table below shows the distribution of patients with good outcomes at 6-month follow-up.

		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?

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.

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.

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.

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?

Generalizing the results

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

- Note that patients in this study had specific characteristics and volunteered to participate;

Generalizing the results

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

- Note that patients in this study had specific characteristics and volunteered to participate;
- They may not be representative of all patients with chronic fatigue syndrome.

Generalizing the results

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

- Note that patients in this study had specific characteristics and volunteered to participate;
- They may not be representative of all patients with chronic fatigue syndrome.
- Results of 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.

To summarize...

Two main data problems:

1. Observed differences $\stackrel{?}{=}$ real differences
   (central topic in following chapters)
2. Results generalizable?
   (this chapter)

Data basics

Our “entry survey”

We conducted an “entry survey” on all of you before class begins. Below are the questions on the survey, and the corresponding variables the response data were stored in:

- `gender`: What is your gender?
- `software_before`: Have you ever used statistical computing softwares (e.g., R, SAS, SPSS, etc.) before?
- `mean_estimate`: Suppose we have observed the following quantities: ... (data omitted). What do you estimate as the mean (or average) of these quantities?
- `courses`: How many online classes have you taken before?
- `confidence`: On a scale from 1 to 5, how confident are you about passing STA 101?

Data matrix

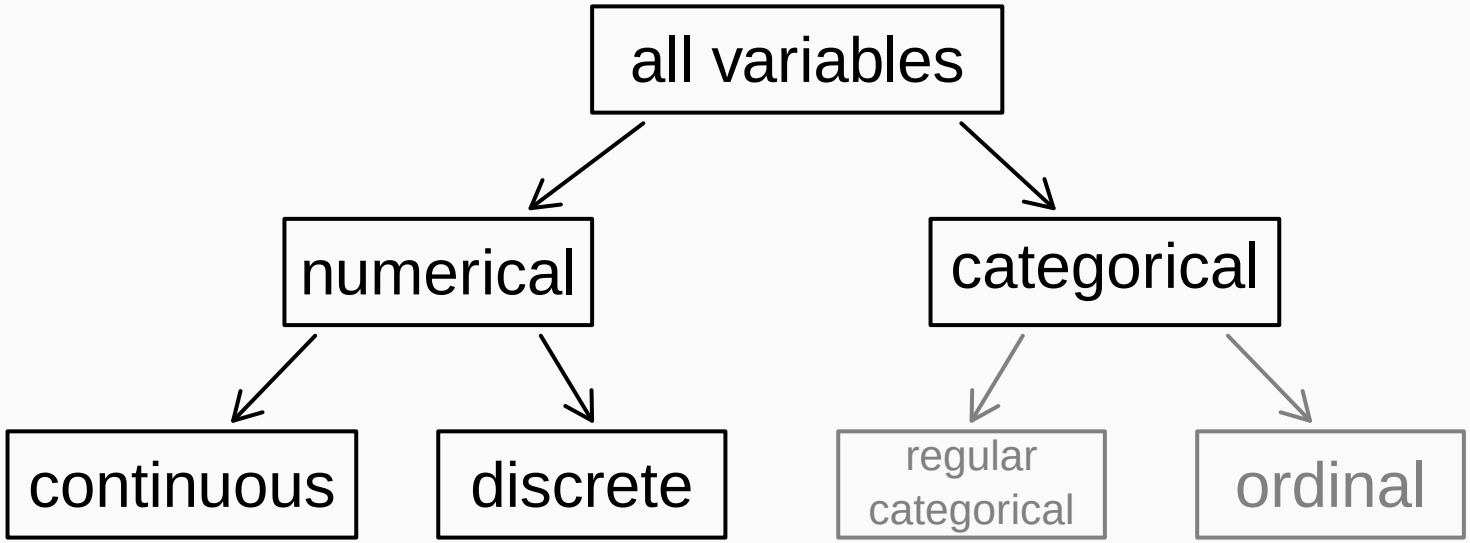
Data collected in the survey can be organized into a data matrix (or data table):

variable ↓

Stu.	gender	software_before	···	confidence
1	Female	No	···	4
2	Male	No	···	4
3	Female	Yes	···	3
4	Female	No	···	5
⋮	⋮	⋮	⋮	⋮
12	Female	Yes	···	4

← *observation*

Types of variables



Types of variables (cont.)

	gender	software_before	mean_estimate	courses	confidence
1	Female	No	3.69	5	4
2	Male	No	4.80	8	4
3	Female	Yes	3.69	4	3
4	Female	No	3.69	7	5
5	Female	No	5.5	10	4
6	Male	No	3.69	8	4

- gender:

Types of variables (cont.)

	gender	software_before	mean_estimate	courses	confidence
1	Female	No	3.69	5	4
2	Male	No	4.80	8	4
3	Female	Yes	3.69	4	3
4	Female	No	3.69	7	5
5	Female	No	5.5	10	4
6	Male	No	3.69	8	4

- `gender`: *categorical*

Types of variables (cont.)

	gender	software_before	mean_estimate	courses	confidence
1	Female	No	3.69	5	4
2	Male	No	4.80	8	4
3	Female	Yes	3.69	4	3
4	Female	No	3.69	7	5
5	Female	No	5.5	10	4
6	Male	No	3.69	8	4

- gender: *categorical*
- software_before:

Types of variables (cont.)

	gender	software_before	mean_estimate	courses	confidence
1	Female	No	3.69	5	4
2	Male	No	4.80	8	4
3	Female	Yes	3.69	4	3
4	Female	No	3.69	7	5
5	Female	No	5.5	10	4
6	Male	No	3.69	8	4

- `gender`: *categorical*
- `software_before`: *categorical*

Types of variables (cont.)

	gender	software_before	mean_estimate	courses	confidence
1	Female	No	3.69	5	4
2	Male	No	4.80	8	4
3	Female	Yes	3.69	4	3
4	Female	No	3.69	7	5
5	Female	No	5.5	10	4
6	Male	No	3.69	8	4

- gender: *categorical*
- software_before: *categorical*
- mean_estimate:

Types of variables (cont.)

	gender	software_before	mean_estimate	courses	confidence
1	Female	No	3.69	5	4
2	Male	No	4.80	8	4
3	Female	Yes	3.69	4	3
4	Female	No	3.69	7	5
5	Female	No	5.5	10	4
6	Male	No	3.69	8	4

- `gender`: *categorical*
- `software_before`: *categorical*
- `mean_estimate`: *numerical, continuous*

Types of variables (cont.)

	gender	software_before	mean_estimate	courses	confidence
1	Female	No	3.69	5	4
2	Male	No	4.80	8	4
3	Female	Yes	3.69	4	3
4	Female	No	3.69	7	5
5	Female	No	5.5	10	4
6	Male	No	3.69	8	4

- `gender`: *categorical*
- `software_before`: *categorical*
- `mean_estimate`: *numerical, continuous*
- `courses`:

Types of variables (cont.)

	gender	software_before	mean_estimate	courses	confidence
1	Female	No	3.69	5	4
2	Male	No	4.80	8	4
3	Female	Yes	3.69	4	3
4	Female	No	3.69	7	5
5	Female	No	5.5	10	4
6	Male	No	3.69	8	4

- `gender`: *categorical*
- `software_before`: *categorical*
- `mean_estimate`: *numerical, continuous*
- `courses`: *numerical, discrete*

Types of variables (cont.)

	gender	software_before	mean_estimate	courses	confidence
1	Female	No	3.69	5	4
2	Male	No	4.80	8	4
3	Female	Yes	3.69	4	3
4	Female	No	3.69	7	5
5	Female	No	5.5	10	4
6	Male	No	3.69	8	4

- `gender`: *categorical*
- `software_before`: *categorical*
- `mean_estimate`: *numerical, continuous*
- `courses`: *numerical, discrete*
- `confidence`:

Types of variables (cont.)

	gender	software_before	mean_estimate	courses	confidence
1	Female	No	3.69	5	4
2	Male	No	4.80	8	4
3	Female	Yes	3.69	4	3
4	Female	No	3.69	7	5
5	Female	No	5.5	10	4
6	Male	No	3.69	8	4

- `gender`: *categorical*
- `software_before`: *categorical*
- `mean_estimate`: *numerical, continuous*
- `courses`: *numerical, discrete*
- `confidence`: *categorical, ordinal - could also be used as numerical*

Practice

What type of variable is a telephone area code?

E.g., 919 (Raleigh-Durham), 310 (Beverly Hills).

(a) numerical, continuous

(b) numerical, discrete

(c) categorical

(d) categorical, ordinal

Practice

What type of variable is a telephone area code?

E.g., 919 (Raleigh-Durham), 310 (Beverly Hills).

(a) numerical, continuous

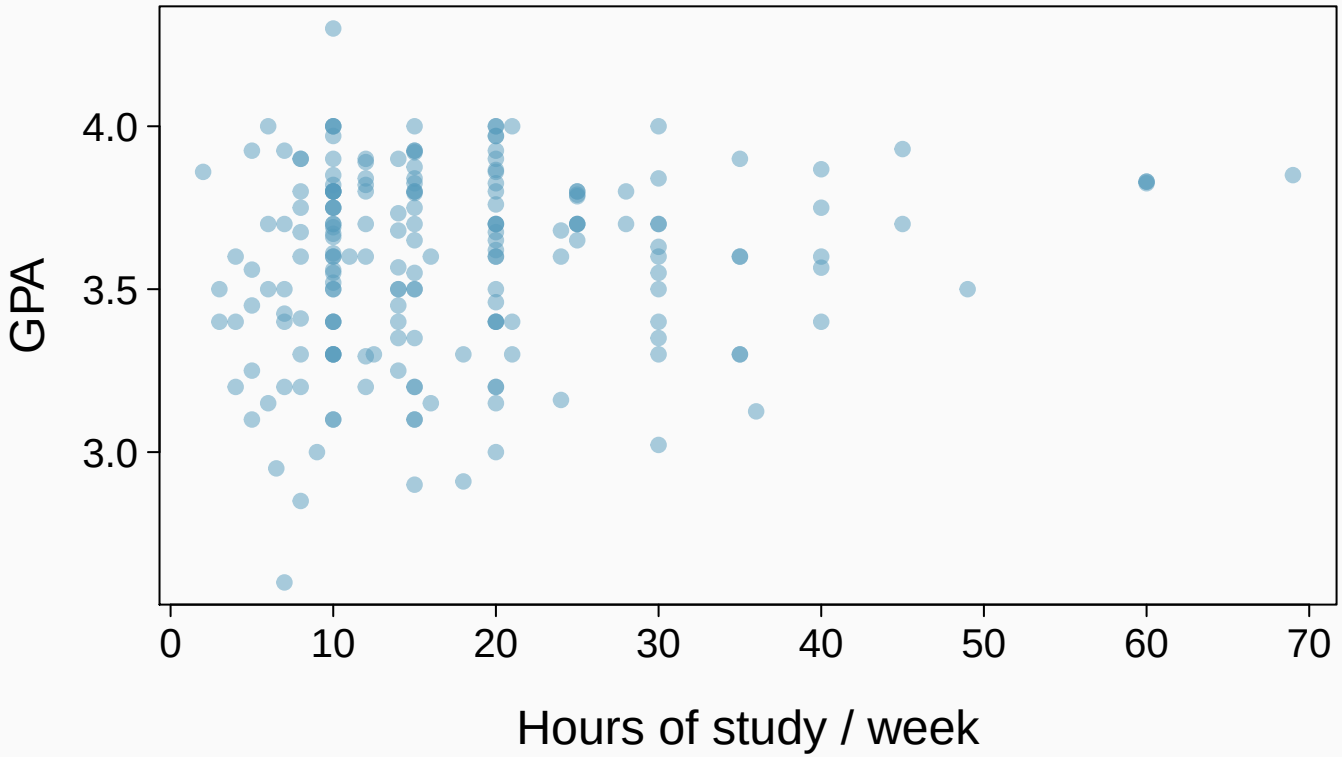
(b) numerical, discrete

(c) *categorical*

(d) categorical, ordinal

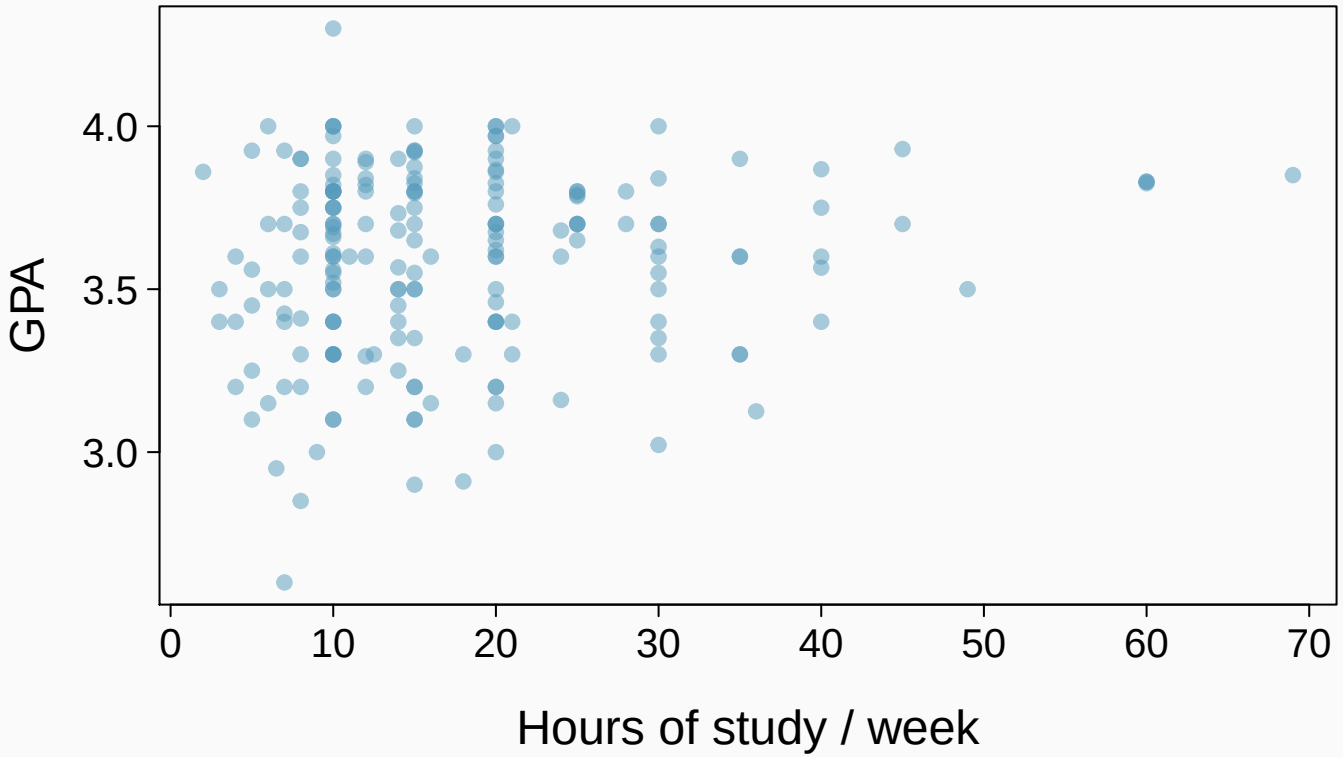
Relationships among variables

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



Relationships among variables

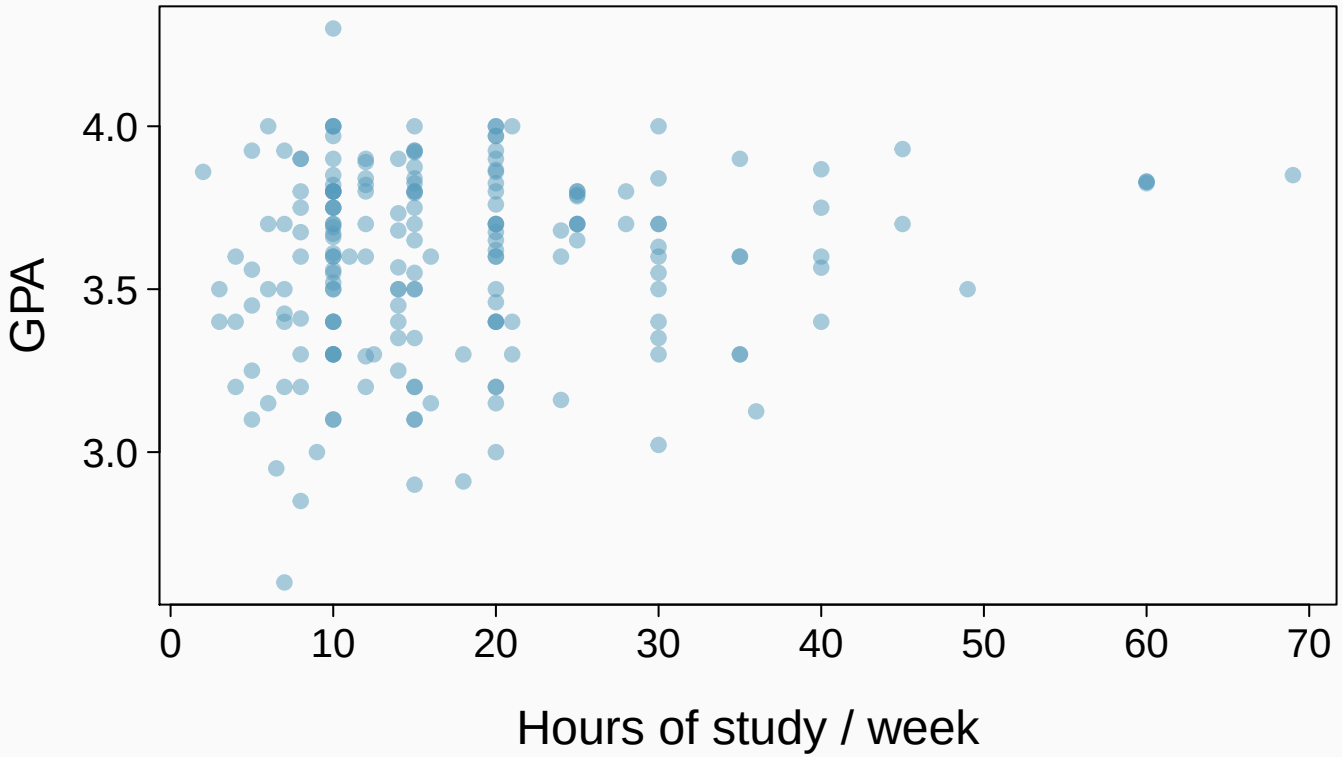
Does there appear to be a relationship between GPA and number of hours students study per week?



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

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.

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.

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.
- *Association ≠ Causation*

Two primary types of data collection

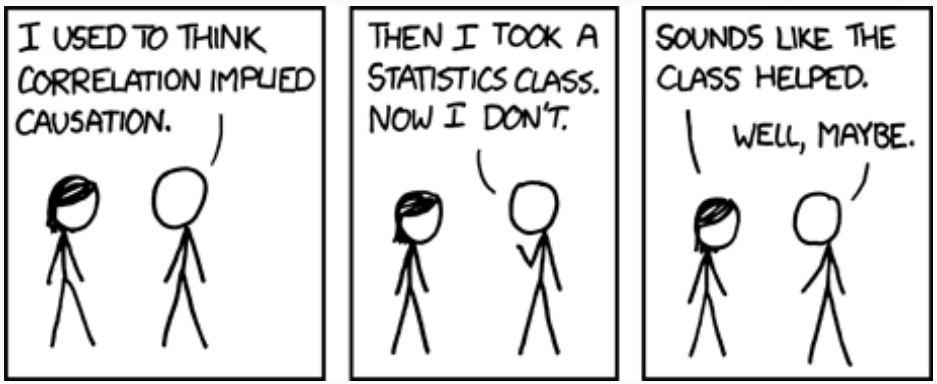
- *Observational studies:* Collect data in a way that does not directly interfere with how the data arise (e.g. surveys).
    - Can provide evidence of a naturally occurring association between variables, but they cannot by themselves show a causal connection.

Two primary types of data collection

- *Observational studies:* Collect data in a way that does not directly interfere with how the data arise (e.g. surveys).
    - Can provide evidence of a naturally occurring association between variables, but they cannot by themselves show a causal connection.
- *Experiment:* Researchers randomly assign subjects to various treatments in order to establish causal connections between the explanatory and response variables.

Association vs. causation

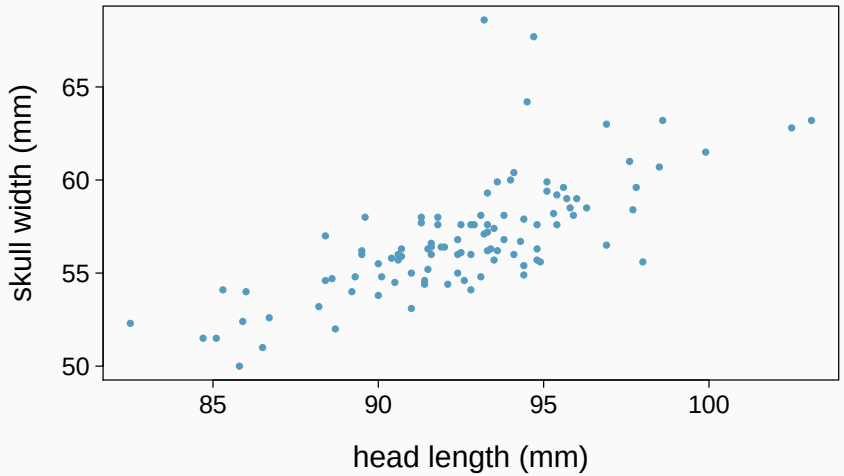
- When two variables show some connection with one another, they are called *associated* variables.
  - Associated variables can also be called *dependent* variables and vice-versa.
- If two variables are not associated, i.e. there is no evident connection between the two, then they are said to be *independent*.
- In general, association does not imply causation, and causation can only be inferred from a *randomized experiment*.



http://xkcd.com/552/

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.

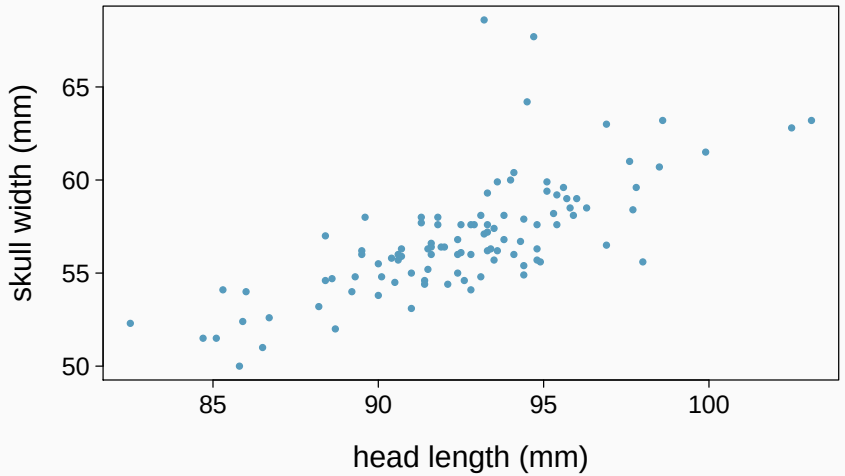
(c) Skull width and head length are negatively associated.

(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.

Sampling principles and strategies

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?

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:

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*

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

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:

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. Some smokers seemed to be sensitive to cigarette smoke, others were completely unaffected.

Anecdotal evidence and early smoking research

- Anti-smoking research started in the 1930s and 1940s when cigarette smoking became increasingly popular. Some smokers seemed to be sensitive to cigarette smoke, others were completely unaffected.
- Anti-smoking research was faced with resistance based on *anecdotal evidence* like “My uncle smokes three packs a day and he’s in perfectly good health”, evidence based on a limited sample size that may not be representative of the population.

Anecdotal evidence and early smoking research

- Anti-smoking research started in the 1930s and 1940s when cigarette smoking became increasingly popular. Some smokers seemed to be sensitive to cigarette smoke, others were completely unaffected.
- Anti-smoking research was faced with resistance based on *anecdotal evidence* like “My uncle smokes three packs a day and he’s in perfectly good health”, evidence based on a limited sample size that may not be representative of the population.
- Concluded then that “smoking is a complex human behavior, by its nature difficult to study, confounded by human variability.”

Anecdotal evidence and early smoking research

- Anti-smoking research started in the 1930s and 1940s when cigarette smoking became increasingly popular. Some smokers seemed to be sensitive to cigarette smoke, others were completely unaffected.
- Anti-smoking research was faced with resistance based on *anecdotal evidence* like “My uncle smokes three packs a day and he’s in perfectly good health”, evidence based on a limited sample size that may not be representative of the population.
- Concluded then 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), it became much clearer in the trends that smoking has negative health impacts.

Census

- Wouldn't it be better to just include everyone and "sample" the entire population?
  - This is called a *census*.

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.

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.

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*.

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*.

Exploratory analysis to 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.

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.

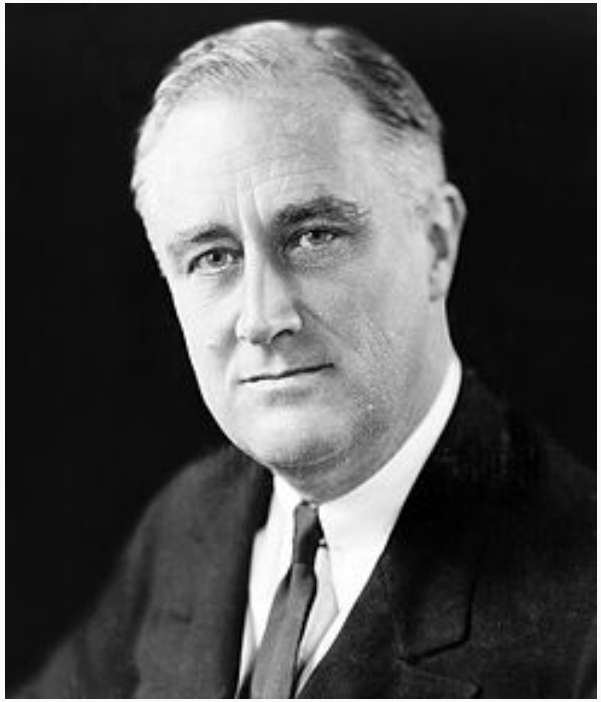
Sampling bias

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

Sampling bias example: Landon vs. FDR

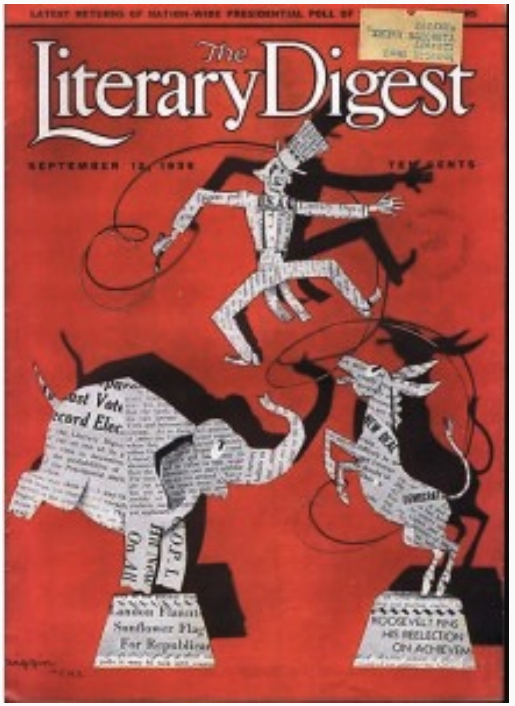
A historical example of a biased sample yielding misleading results:

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.

Observational studies

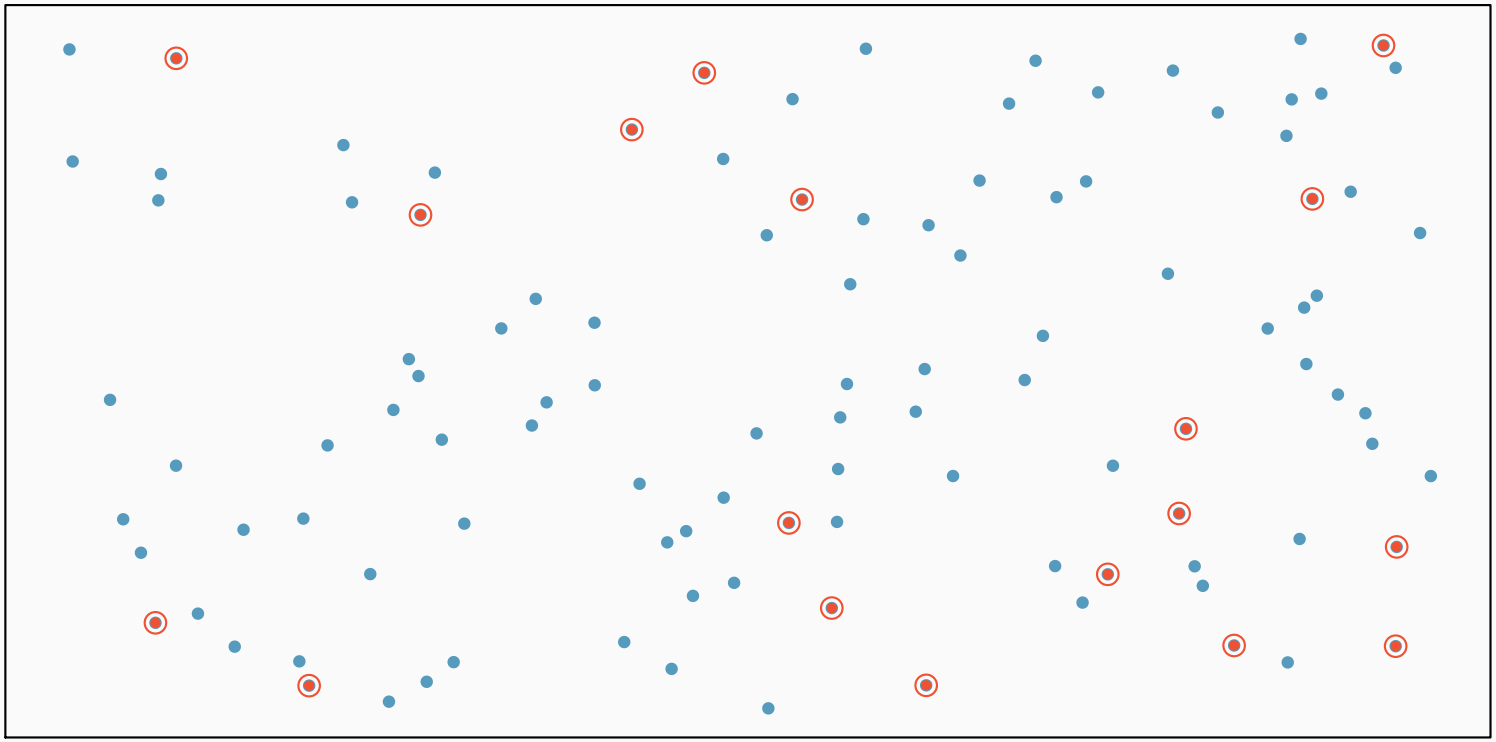
- Researchers collect data in a way that does not directly interfere with how the data arise.
- Results of an observational study can generally be used to establish an association between the explanatory and response variables.

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.

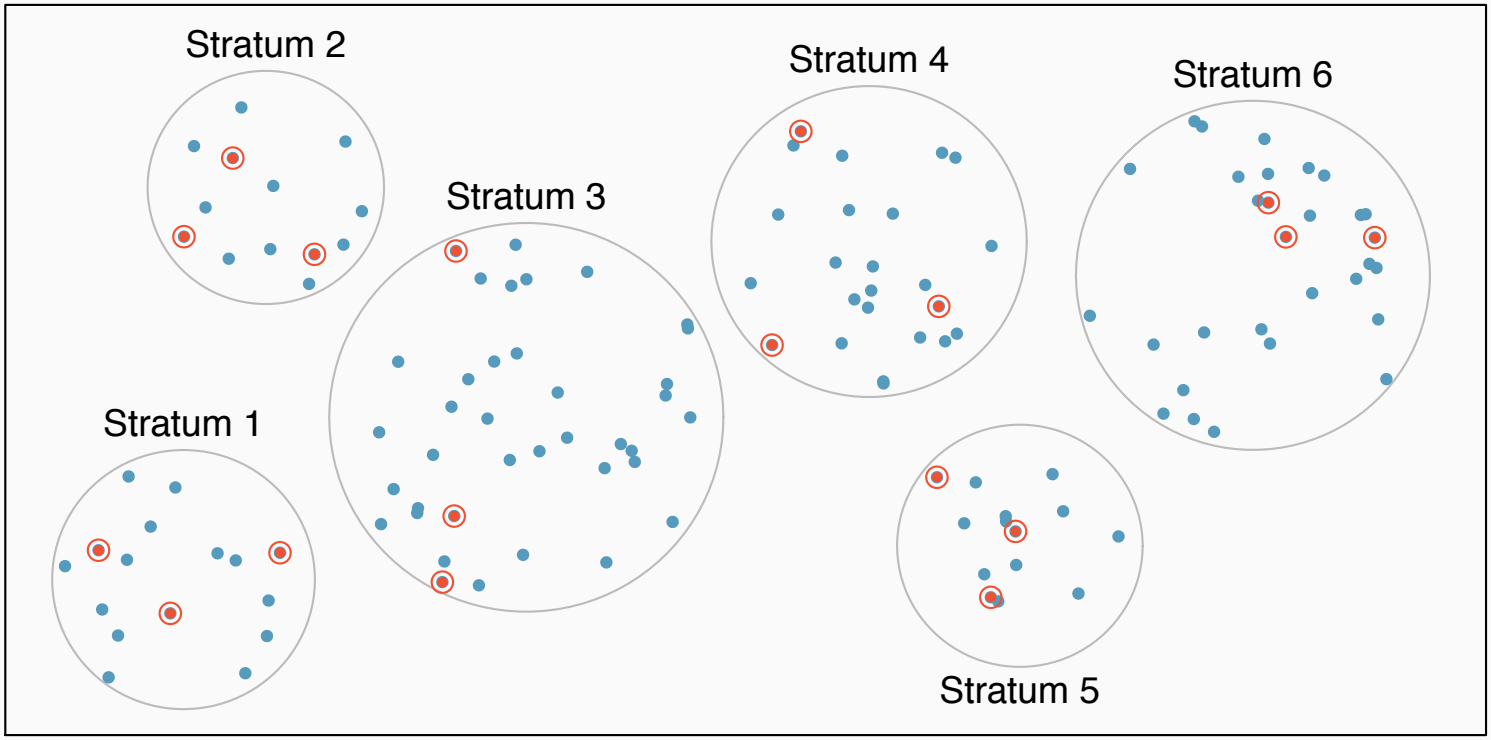
Simple random sample

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



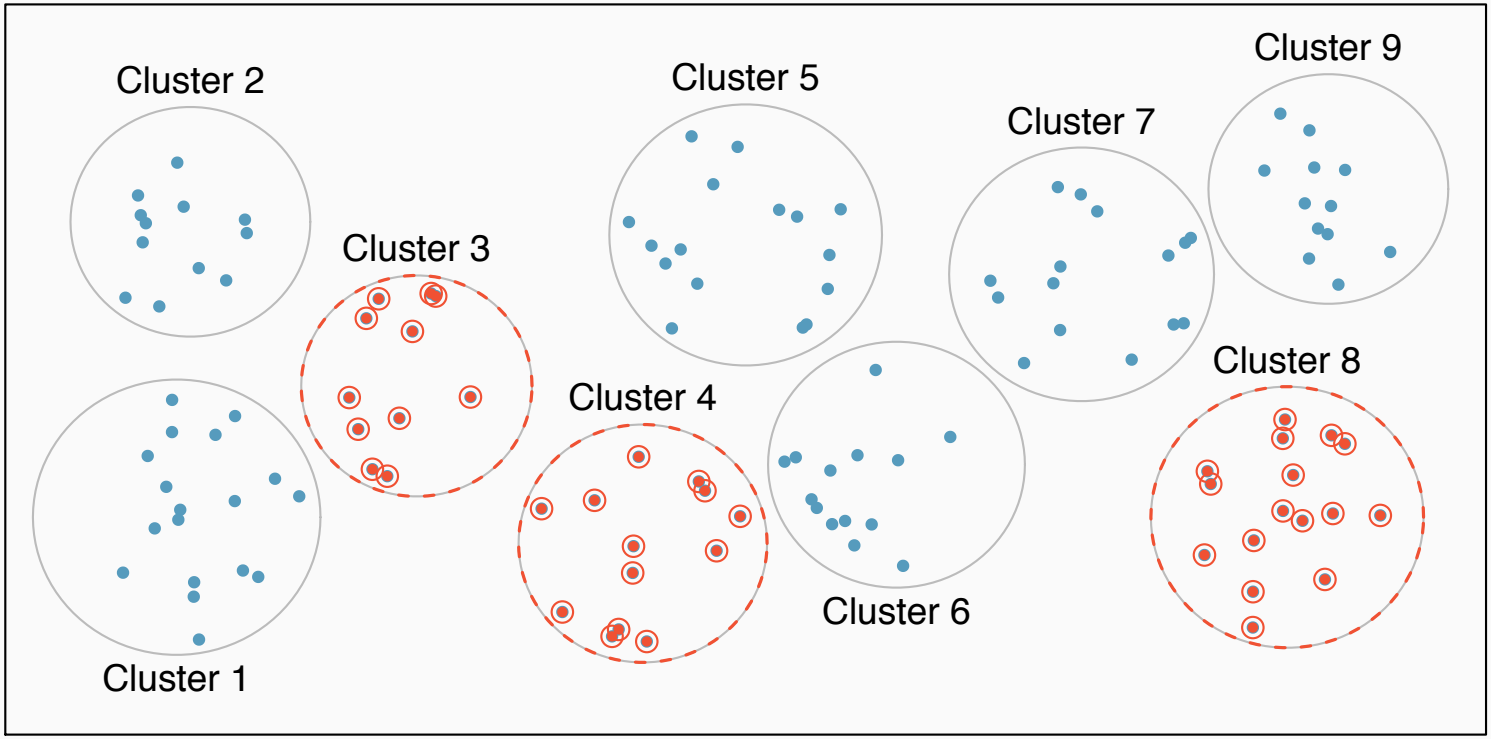
Stratified sample

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



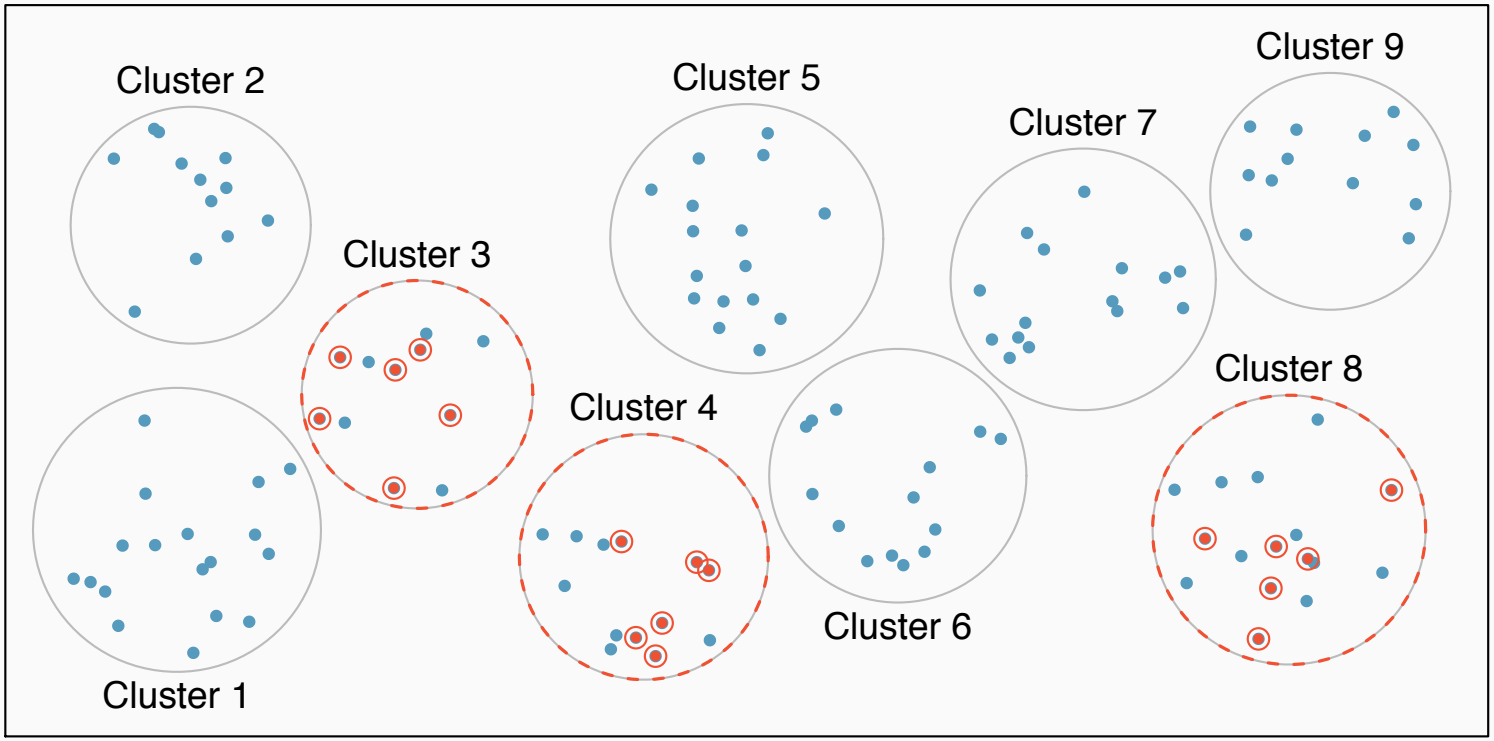
Cluster sample

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:* Control for the (potential) effect of variables other than the ones directly being studied.
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 gel makes you run faster:

More on blocking

- We would like to design an experiment to investigate if energy gel makes you run faster:
  - *Treatment*: energy gel
  - *Control*: no energy gel

More on blocking

- We would like to design an experiment to investigate if energy gel makes you run faster:
  - *Treatment*: energy gel
  - *Control*: no energy gel
- It is suspected that energy gel might affect *pro* and *amateur* athletes differently, therefore we block for pro status:

More on blocking

- We would like to design an experiment to investigate if energy gel makes you run faster:
  - *Treatment*: energy gel
  - *Control*: no energy gel
- It is suspected that energy gel 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

More on blocking

- We would like to design an experiment to investigate if energy gel makes you run faster:
  - *Treatment*: energy gel
  - *Control*: no energy gel
- It is suspected that energy gel 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

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.
- 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

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