Announcements

- Homework
  - 1st one will be posted this afternoon
  - Due next Monday night
- Wednesday In-class Lab
  - Instructions for getting R and RStudio installed are on webpage
  - Bring your computer with R and RStudio on it Wednesday!
- Read through the rest of Chapter 1 by Friday

Warm Up

What type of variable is a telephone area code?

A) numerical, continuous

B) numerical, discrete

C) categorical

D) categorical, ordinal

Warm Up

What type of variable is a telephone area code?

A) numerical, continuous

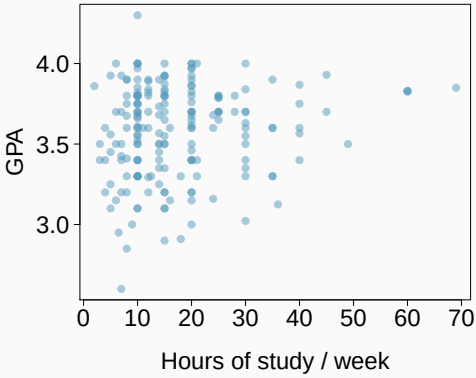
B) numerical, discrete

C) categorical

D) categorical, ordinal

Relationships among variables

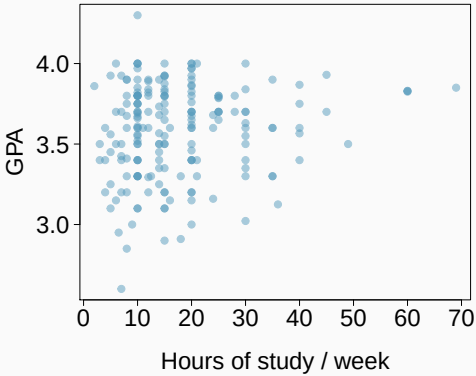
Does there appear to be a relationship between GPA and number of hours students study per week?



Relationships among variables

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

- At the very least low scores seem to improve as study hours increase

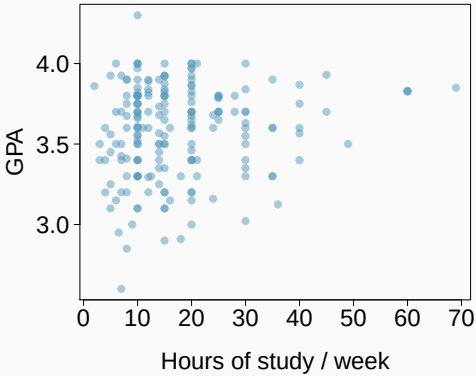


Relationships among variables

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

- At the very least low scores seem to improve as study hours increase

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

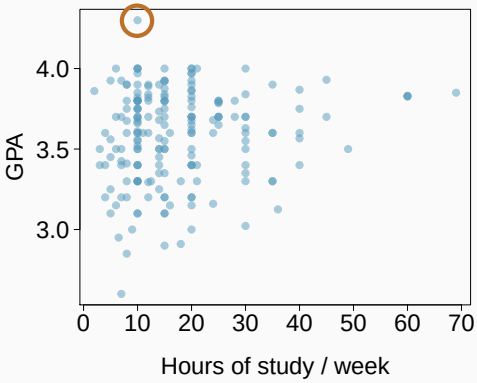


Relationships among variables

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

- At the very least low scores seem to improve as study hours increase

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



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***.

Practice

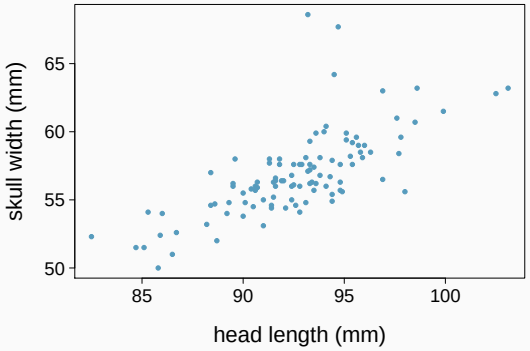
Based on the scatterplot below, which of the following statements is correct about the head and skull lengths of possums?

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.



Practice

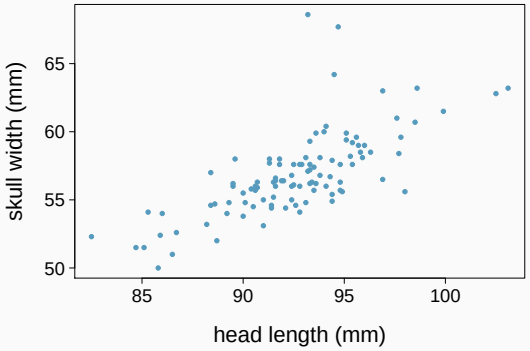
Based on the scatterplot below, which of the following statements is correct about the head and skull lengths of possums?

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.



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

Source: here

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

Source: here

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

Source: here

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

Source: here

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

Source: here

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

Source: here

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?
  - Sure! This is called a ***census***.

Census

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

NATIONAL

Census Bureau To Test How Controversial Citizenship Question Affects Responses

December 6, 2018 · 2:40 PM ET

HANSI LO WANG

Newly sworn-in U.S. citizens stand during a naturalization ceremony in Alexandria, Va., in August. The Census Bureau is planning to test how a question about U.S. citizenship status the Trump administration added will affect responses to the 2020 census.

Claire Harbage/NPR

Source: here

Exploratory analysis to inference

- Sampling is natural.

Exploratory analysis to inference

- Sampling is natural.
- Imagine 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.
- Imagine 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.
- Imagine 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

- Sampling is natural.
- Imagine 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.

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.

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

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

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

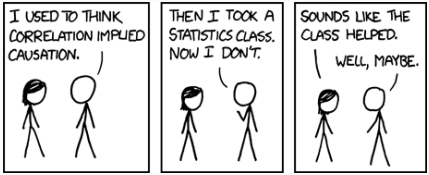
- ***Observational study:*** Researchers collect data in a way that does not directly interfere with how the data arise, i.e. they merely "observe", and can only establish an association between the explanatory and response variables.

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 nothing else sticks from this class, at least remember that "correlation does not imply causation".



http://xkcd.com/552/

Observational Studies and Sampling Strategies

Study: Cereal Keeps Girls Slim

BY CHRISTINE LAGORIO
UPDATED ON: SEPTEMBER 9, 2005 / 11:30 AM / CBS/AP

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.

Source: `https://www.cbsnews.com/news/study-cereal-keeps-girls-slim/`

The Breakdown

- ***Observational Study or Experiment?***

The Breakdown

- ***Observational Study or Experiment?***
  This decidedly an observational study, since researchers merely observed the gir's behavior and didn't force them to eat or not eat breakfast.

The Breakdown

- ***Observational Study or Experiment?***
  This decidedly an observational study, since researchers merely observed the gir's behavior and didn't force them to eat or not eat breakfast.
- ***Conclusions?***

The Breakdown

- ***Observational Study or Experiment?***
  This decidedly an observational study, since researchers merely observed the gir's behavior and didn't force them to eat or not eat breakfast.
- ***Conclusions?***
  There is an association between girls eating breakfast and being slimmer.

The Breakdown

- ***Observational Study or Experiment?***
  This decidedly an observational study, since researchers merely observed the gir's behavior and didn't force them to eat or not eat breakfast.
- ***Conclusions?***
  There is an association between girls eating breakfast and being slimmer.
- ***Who sponsored the study?***

The Breakdown

- ***Observational Study or Experiment?***
  This decidedly an observational study, since researchers merely observed the gir's behavior and didn't force them to eat or not eat breakfast.
- ***Conclusions?***
  There is an association between girls eating breakfast and being slimmer.
- ***Who sponsored the study?***
  General Mills.

3 possible explanations

3 possible explanations

1. Eating breakfast causes girls to be thinner.

3 possible explanations

1. Eating breakfast causes girls to be thinner.

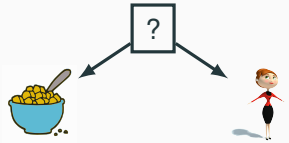
2. Being thin causes girls to eat breakfast.

3 possible explanations

1. Eating breakfast causes girls to be thinner.

2. Being thin causes girls to eat breakfast.

3. A third variable causes both!

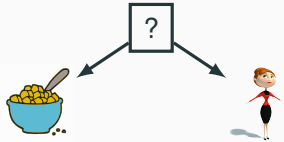


3 possible explanations

1. Eating breakfast causes girls to be thinner.

2. Being thin causes girls to eat breakfast.

3. A third variable causes both!



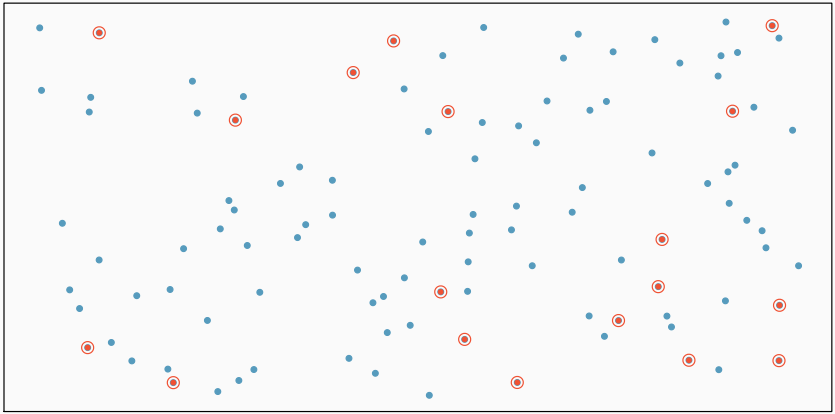
An extraneous variable that affects both explanatory and the response variable and make it seem like there is a relationship between the two are call ***confounding*** variables

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.

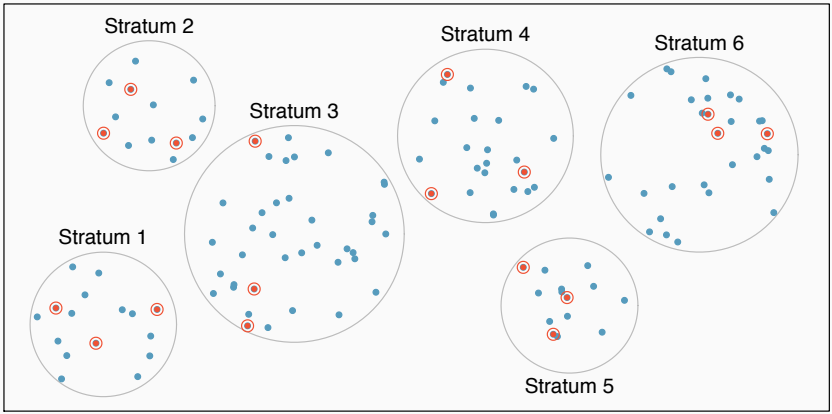
Simple random sample

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



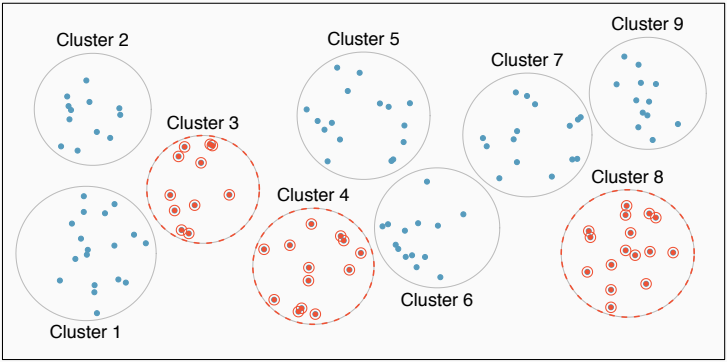
Stratified sample

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

Multistage samples are a bit of a combination. We take a simple random sample of clusters, and then take a simple random sample of observations from the sampled clusters.