Introduction

Portland State University
USP 634 Data Analysis I
Spring 2018

What is Statistics?

“Statistics is the study of the collection, analysis, interpretation, presentation, and organization of data”. -- Wikipedia

- **descriptive statistics** - summarizes data from a sample using indexes such as the mean or standard deviation
- **inferential statistics -** draws conclusions from data that are subject to random variation

Why Statistics?

“Statistics ... the most important science in the whole world: for upon it depends the practical application of every other science and of every art; the one science essential to all political and social administration, all education, all organization based upon experience, for it only gives the results of our experience.” -- Florence Nightingale

Davidian, M., and Louis, T.A., 2012. Why Statistics?, Science, Vol. 336, Issue 6077, pp. 12.
Thomas H. Davenport and D.J. Patil, 2012, Data Scientist: The Sexiest Job of the 21st Century, Harvard Business Review, October 2012, page 70-76.

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.

Sources: https://www.cbsnews.com/news/study-cereal-keeps-girls-slim/
Barton, Bruce A. et al., 2005, The Relationship of Breakfast and Cereal Consumption to Nutrient Intake and Body Mass Index: The National Heart, Lung, and Blood Institute Growth and Health Study, Journal of the American Dietetic Association , Volume 105 , Issue 9 , 1383 - 1389

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?

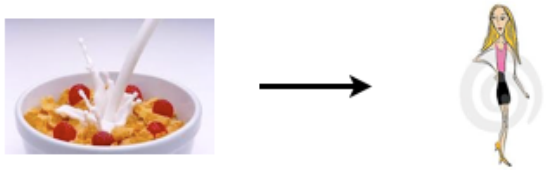
There is an association between girls eating breakfast and being slimmer.

Who sponsored the study?

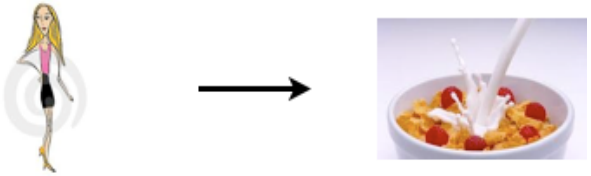
General Mills.

3 Possible Explanations

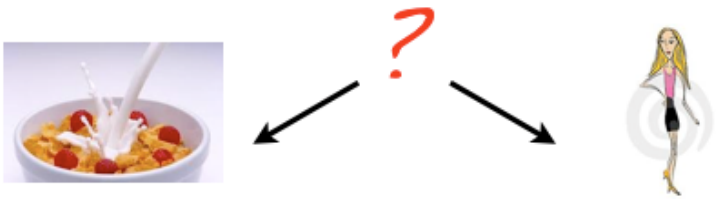
1. Eating breakfast causes girls to be thinner.



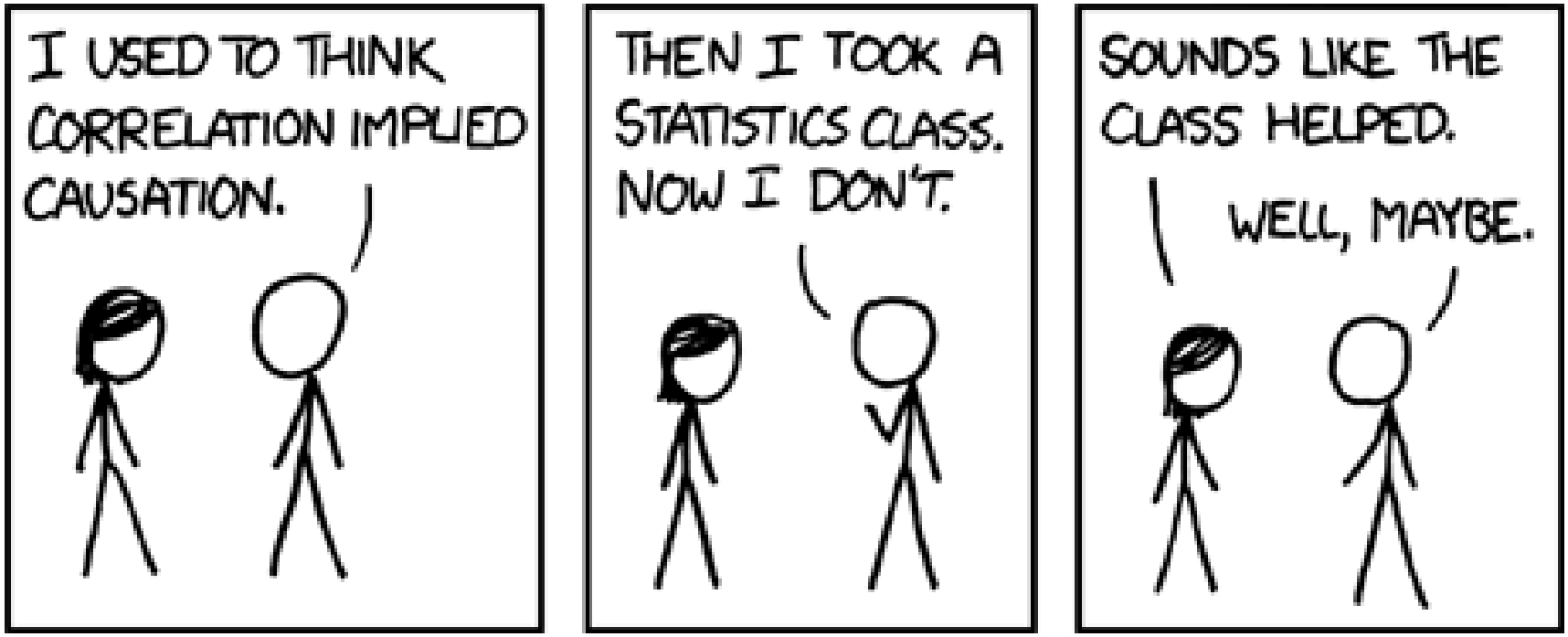
2. Being thin causes girls to eat breakfast.



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.



Correlation vs Causation



Source: http://xkcd.com/552

Anecdotal evidence and early smoking research

- Anti-smoking research started in the 1930s and 1940s when cigarette smoking became increasingly popular. While some smokers seemed to be sensitive to cigarette smoke, others were completely unaffected.

- Anti-smoking research was faced with resistance based on anecdotal evidence such as "My uncle smokes three packs a day and he's in perfectly good health", evidence based on a limited sample size that might not be representative of the population.

- It was concluded that "smoking is a complex human behavior, by its nature difficult to study, confounded by human variability."

- In time researchers were able to examine larger samples of cases (smokers), and trends showing that smoking has negative health impacts became much clearer.

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

Census

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

- There are problems with taking a census:
  - It can be difficult to complete a census: there always seem to be some individuals who are hard to locate or hard to measure. *And these difficult-to-find people may have certain characteristics that distinguish them from the rest of the population.*
  - Populations rarely stand still. Even if you could take a census, the population changes constantly, so it's never possible to get a perfect measure.
  - Taking a census may be more complex than sampling.

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.

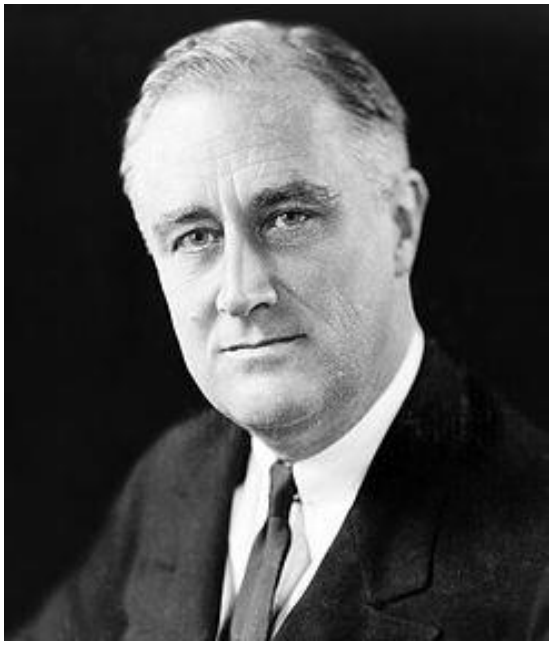
Voluntary response: Occurs when the sample consists of people who volunteer to respond because they have strong opinions on the issue. Such a sample will also not be representative of the population.

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

Sampling bias example: Landon vs. FDR

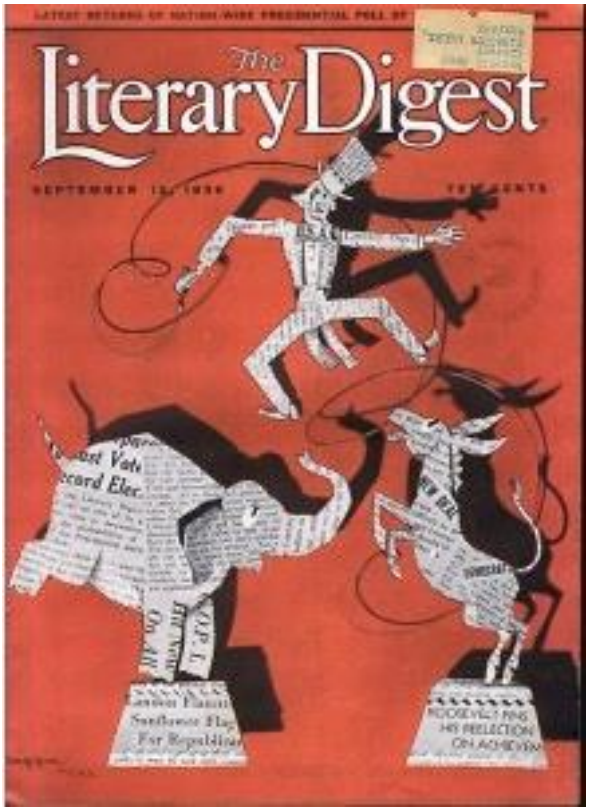
A historical example of a biased sample yielding misleading results

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:

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

- 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

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.

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.

Type I error and Type II error

		Truth	
		Not Guilty	Guilty
Verdict	Guilty	Type I Error -- Innocent person goes to jail (and maybe guilty person goes free)	Correct Decision
	Not Guilty	Correct Decision	Type II Error -- Guilty person goes free

Hypothesis: "The evidence produced before the court proves that this man is guilty."
Null hypothesis (H_0): "This man is innocent."

- A correct positive outcome occurs when convicting a guilty person.
- A correct negative outcome occurs when letting an innocent person go free.
- A type I error occurs when convicting an innocent person (a miscarriage of justice). A type II error occurs when letting a guilty person go free (an error of impunity).

Source: https://en.wikipedia.org/wiki/Type_I_and_type_II_errors

Statistics Is a Weird Subject

- It's not like math, it is like math
- It's like a foreign language
- It's like other courses, and there will also be a great deal of practical, conceptual, and other substantive information
- It's progressive - Everything builds on everything else.
- Statistics is a unique and unusual topic involving some very abstract and weird ideas, but is extremely useful tools.

Credit: Dr. Jason Newsom

Data Basics

Data matrix

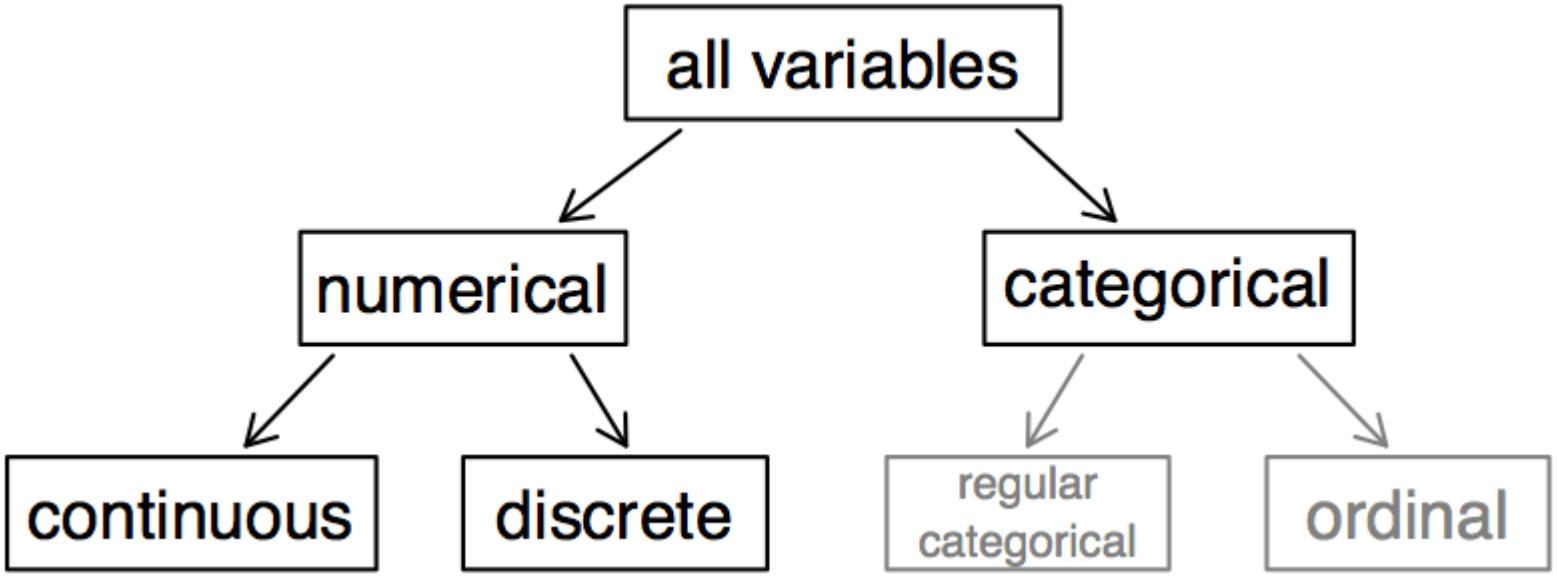
Data collected on students in a statistics class on a variety of variables:

variable ↓

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

← *observation*

Types of variables



Types of variables (cont.)

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

gender - 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

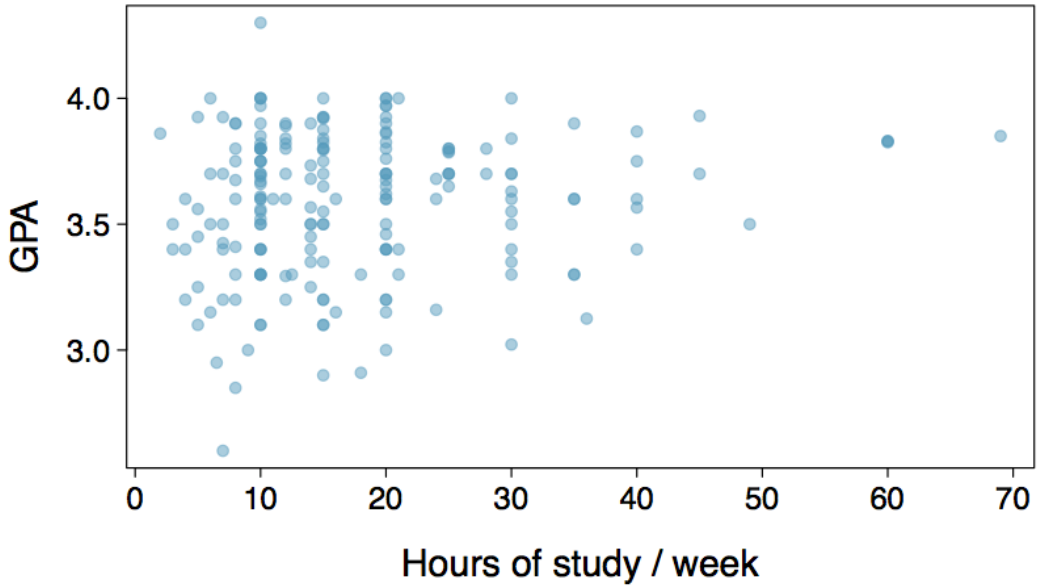
Practice

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 the hours of study per week and the GPA of a student?



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

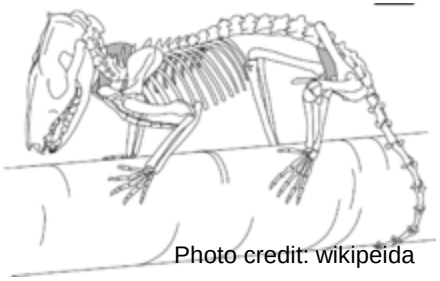
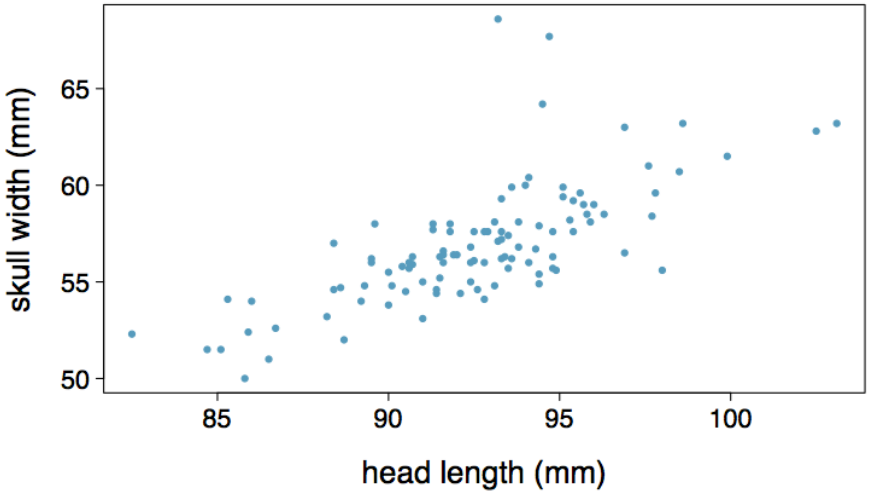


Photo credit: wikipeida

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



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

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

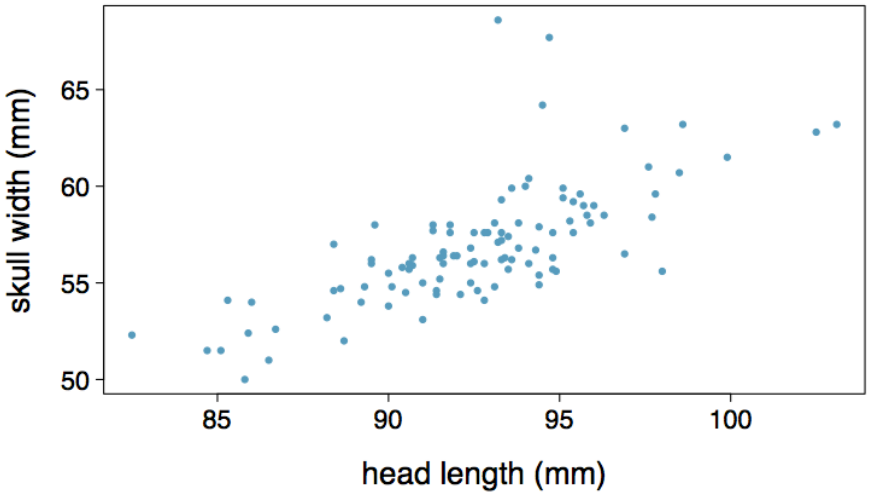
(c) Skull width and head length are negatively associated.

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

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

Practice

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



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

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

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

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

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

Associated vs. independent

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