Section 001
CDS 101 *Introduction to Computational and Data Sciences*

Class 25: Inference and simulations IV / Modeling I

April 24, 2018

General

Annoucements

- Reading for next Tuesday's class: *R for Data Science*
  - From chapter 23: section 23.4 through to the end of section 23.6
  - Last reading!
- Homework 4 due on Friday, April 27th by 11:59pm
- Final project handed out, due on Friday, May 11th by 11:59pm
  - Do not wait to start!

Confidence intervals

Example: Constructing a confidence interval

What is the 95% confidence interval for the *Mythbusters* yawning experiment?

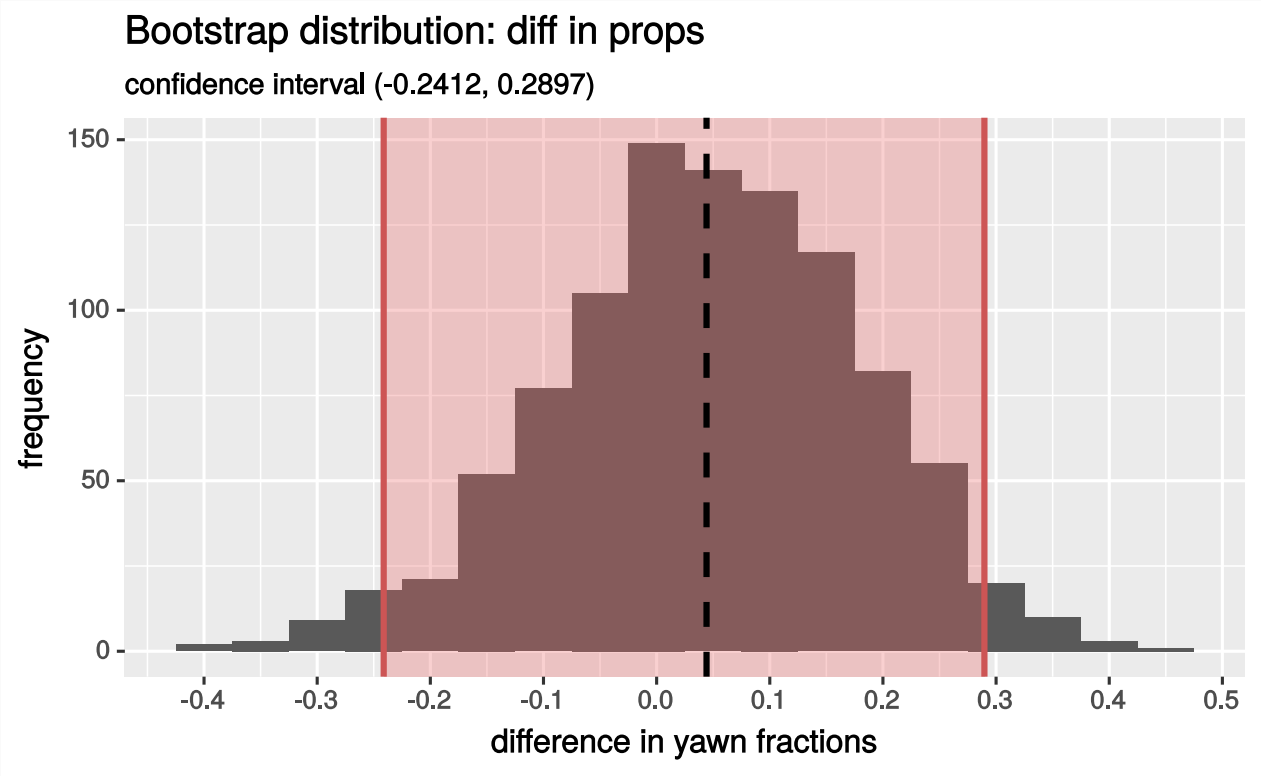
```
yawn_bootstrap <- yawn %>%
  specify(yawn ~ group, success = "yes") %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "diff in props", order = c("Treatment", "Control"))
```

```
yawn_ci_bounds <- yawn_bootstrap %>%
  summarize(
    lower = quantile(stat, probs = c(0.025), type = 1),
    upper = quantile(stat, probs = c(0.975), type = 1))
```

lower	upper
-0.2412281	0.2896825

Example: Constructing a confidence interval

What is the 95% confidence interval for the *Mythbusters* yawning experiment?



People will, on average, yawn 24% less to 29% more when someone near them yawns

What does 95% confident mean?

- Suppose we generated a series of bootstrap distributions with `infer` and found the 95% confidence interval for each one

What does 95% confident mean?

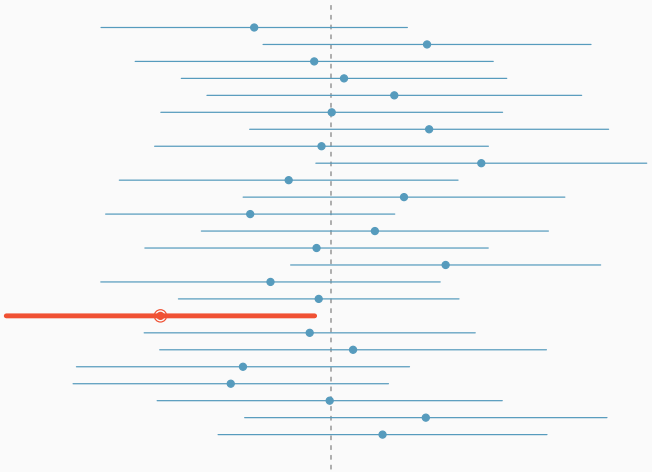
- Suppose we generated a series of bootstrap distributions with `infer` and found the 95% confidence interval for each one
- Let's also assume that, for one reason or another, we somehow knew the value of the true population mean

What does 95% confident mean?

- Suppose we generated a series of bootstrap distributions with `infer` and found the 95% confidence interval for each one
- Let's also assume that, for one reason or another, we somehow knew the value of the true population mean
- Then, the phrase "95% confident" means that about 95% of those intervals would contain the true population mean

What does 95% confident mean?

- Suppose we generated a series of bootstrap distributions with `infer` and found the 95% confidence interval for each one
- Let's also assume that, for one reason or another, we somehow knew the value of the true population mean
- Then, the phrase "95% confident" means that about 95% of those intervals would contain the true population mean
- The figure shows this process with 25 samples, where 24 of the resulting confidence intervals contain the true average number of exclusive relationships, and one does not.



Width of an interval

If we want to be more certain that we capture the population parameter, i.e. increase our confidence level, should we use a wider interval or a smaller interval?

Width of an interval

If we want to be more certain that we capture the population parameter, i.e. increase our confidence level, should we use a wider interval or a smaller interval?

A wider interval.

Width of an interval

If we want to be more certain that we capture the population parameter, i.e. increase our confidence level, should we use a wider interval or a smaller interval?

A wider interval.

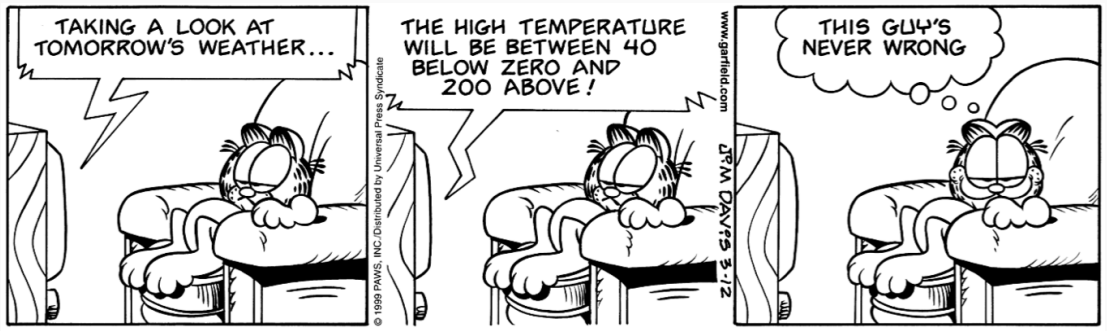
Can you see any drawbacks to using a wider interval?

Width of an interval

If we want to be more certain that we capture the population parameter, i.e. increase our confidence level, should we use a wider interval or a smaller interval?

A wider interval.

Can you see any drawbacks to using a wider interval?



If the interval is too wide it may not be very informative.

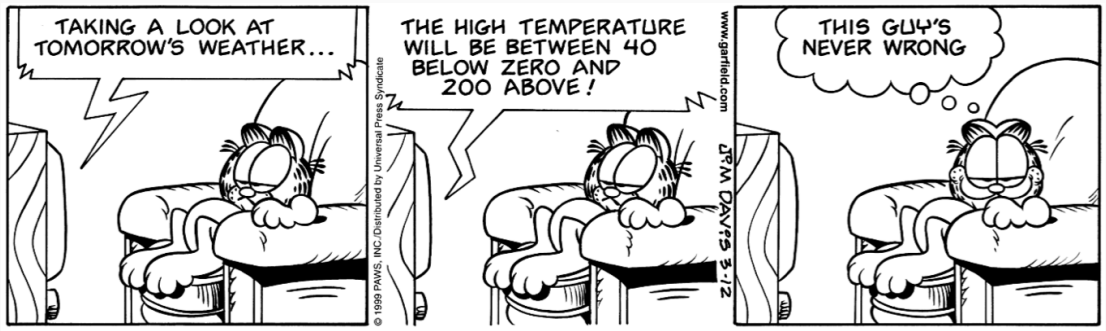
Image source (defunct): http://web.as.uky.edu/statistics/users/earo227/misc/garfield_weather.gif

Width of an interval

If we want to be more certain that we capture the population parameter, i.e. increase our confidence level, should we use a wider interval or a smaller interval?

A wider interval.

Can you see any drawbacks to using a wider interval?



If the interval is too wide it may not be very informative.

- Commonly used confidence levels in practice are 90%, 95%, 98%, and 99%.

Image source (defunct): http://web.as.uky.edu/statistics/users/earo227/misc/garfield_weather.gif

Effect size

- We would like to know how significant a result is, not just whether or not we can reject the null hypothesis.

Effect size

- We would like to know how significant a result is, not just whether or not we can reject the null hypothesis.
- The **effect size** measures the relative difference between two distributions.

Effect size

- We would like to know how significant a result is, not just whether or not we can reject the null hypothesis.
- The **effect size** measures the relative difference between two distributions.
- One effect size definition: Cohen's d

Effect size

- We would like to know how significant a result is, not just whether or not we can reject the null hypothesis.
- The **effect size** measures the relative difference between two distributions.
- One effect size definition: Cohen's *d*

Effect size	d
Very small	0.01
Small	0.20
Medium	0.50
Large	0.80
Very large	1.20
Huge	2.00

Effect size

- We would like to know how significant a result is, not just whether or not we can reject the null hypothesis.
- The **effect size** measures the relative difference between two distributions.
- One effect size definition: Cohen's *d*

Effect size	d
Very small	0.01
Small	0.20
Medium	0.50
Large	0.80
Very large	1.20
Huge	2.00

- Guided instructions on how to use it on Homework 4!

Statistical errors and *p*-hacking

Issues with statistics in modern science

- Over-reliance of *p*-values when determining an experiment's worth

Issues with statistics in modern science

- Over-reliance of p-values when determining an experiment's worth
- Data dredging/p-hacking

Issues with statistics in modern science

- Over-reliance of p-values when determining an experiment's worth
- Data dredging/p-hacking
- Lack of transparency regarding statistical analysis

Issues with statistics in modern science

- Over-reliance of p-values when determining an experiment's worth
- Data dredging/p-hacking
- Lack of transparency regarding statistical analysis
- Poor statistical practices among researchers

Issues with statistics in modern science

- Over-reliance of p-values when determining an experiment's worth
- Data dredging/p-hacking
- Lack of transparency regarding statistical analysis
- Poor statistical practices among researchers
- Lack of reports about experiments that fail to reject the null hypothesis

Issues with statistics in modern science

- Over-reliance of *p*-values when determining an experiment's worth
- Data dredging/*p*-hacking
- Lack of transparency regarding statistical analysis
- Poor statistical practices among researchers
- Lack of reports about experiments that fail to reject the null hypothesis
- Ignoring or underemphasizing effect size

Example: Which political party is better for the economy?

- Start with a reasonable hypothesis: the economy is affected by whether or not Democrats or Republicans are in office

Example: Which political party is better for the economy?

- Start with a reasonable hypothesis: the economy is affected by whether or not Democrats or Republicans are in office
- Collect data about different measures of economic performance and when different politicians were in office

Example: Which political party is better for the economy?

- Start with a reasonable hypothesis: the economy is affected by whether or not Democrats or Republicans are in office
- Collect data about different measures of economic performance and when different politicians were in office
- Construct a basic model connecting the two

Example: Which political party is better for the economy?

- Start with a reasonable hypothesis: the economy is affected by whether or not Democrats or Republicans are in office
- Collect data about different measures of economic performance and when different politicians were in office
- Construct a basic model connecting the two
- FiveThirtyEight Applet (http://53eig.ht/HackingScience)

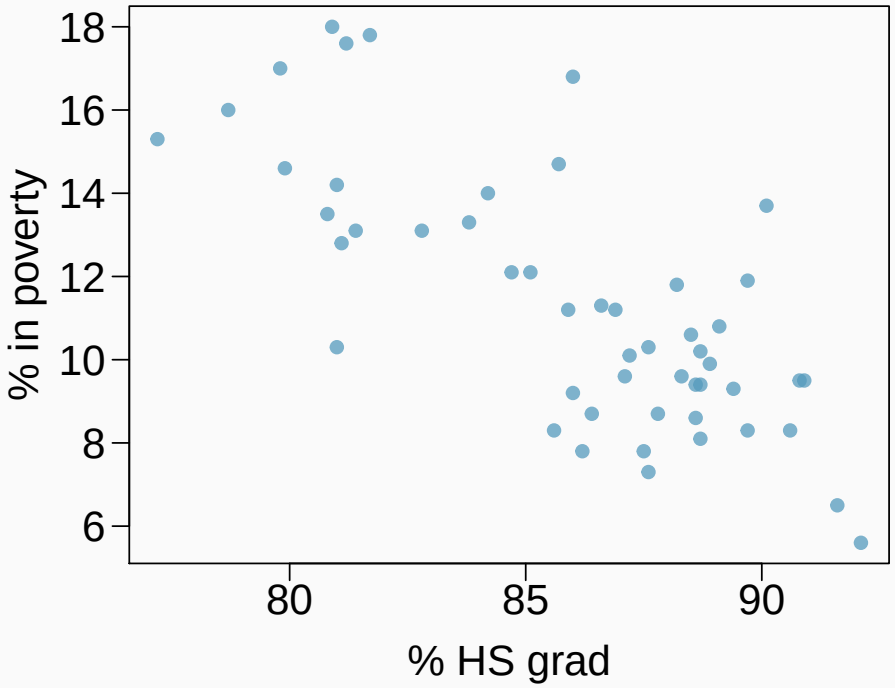
Line fitting, residuals, and correlation

Modeling numerical variables

In this unit we will learn to quantify the relationship between two numerical variables, as well as modeling numerical response variables using a numerical or categorical explanatory variable.

Poverty vs. HS graduate rate

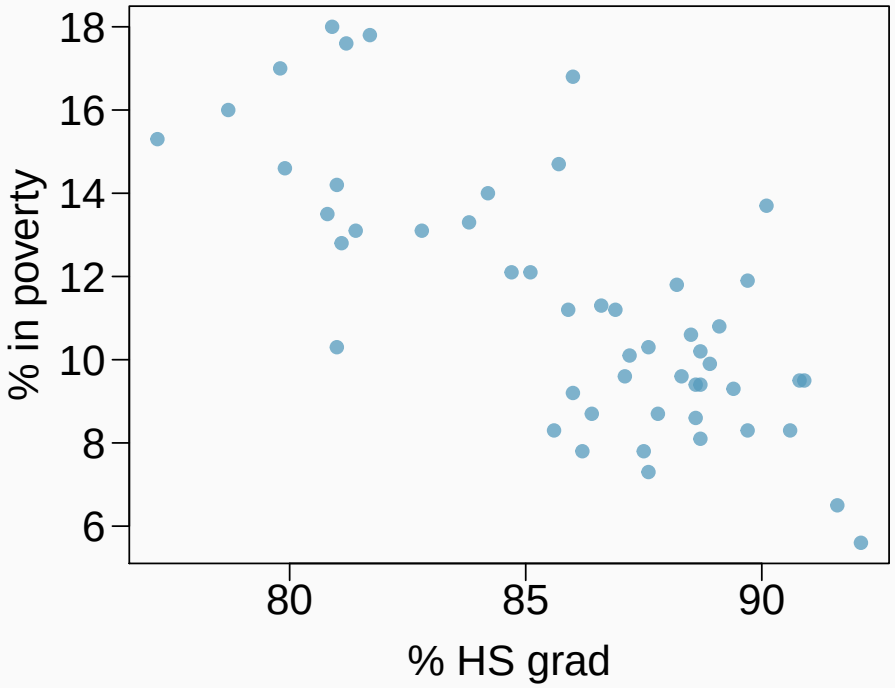
The *scatterplot* below shows the relationship between HS graduate rate in all 50 US states and DC and the % of residents who live below the poverty line (income below $23,050 for a family of 4 in 2012).



Response variable?

Poverty vs. HS graduate rate

The *scatterplot* below shows the relationship between HS graduate rate in all 50 US states and DC and the % of residents who live below the poverty line (income below $23,050 for a family of 4 in 2012).

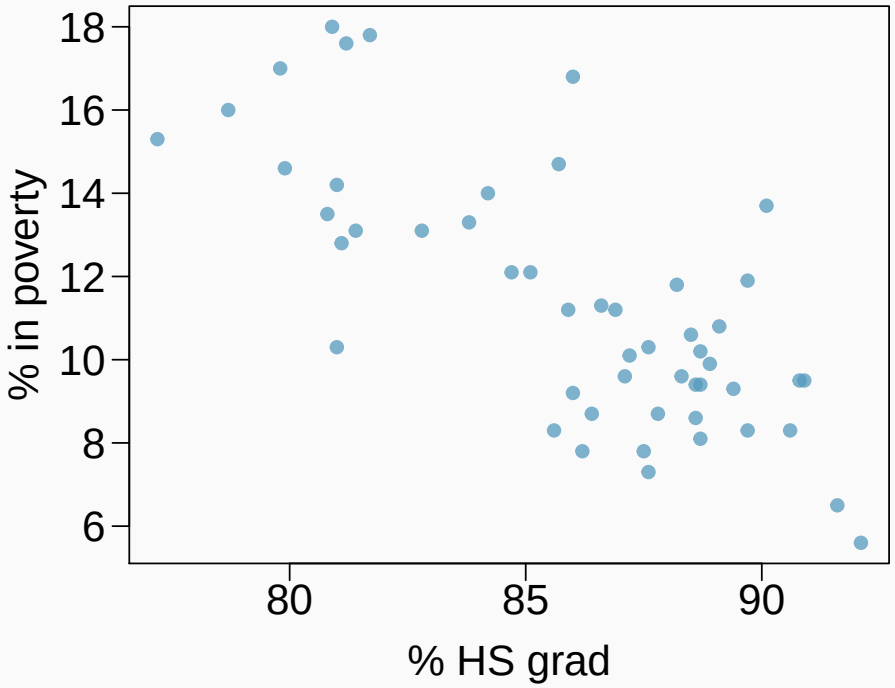


Response variable?

% in poverty

Poverty vs. HS graduate rate

The *scatterplot* below shows the relationship between HS graduate rate in all 50 US states and DC and the % of residents who live below the poverty line (income below $23,050 for a family of 4 in 2012).



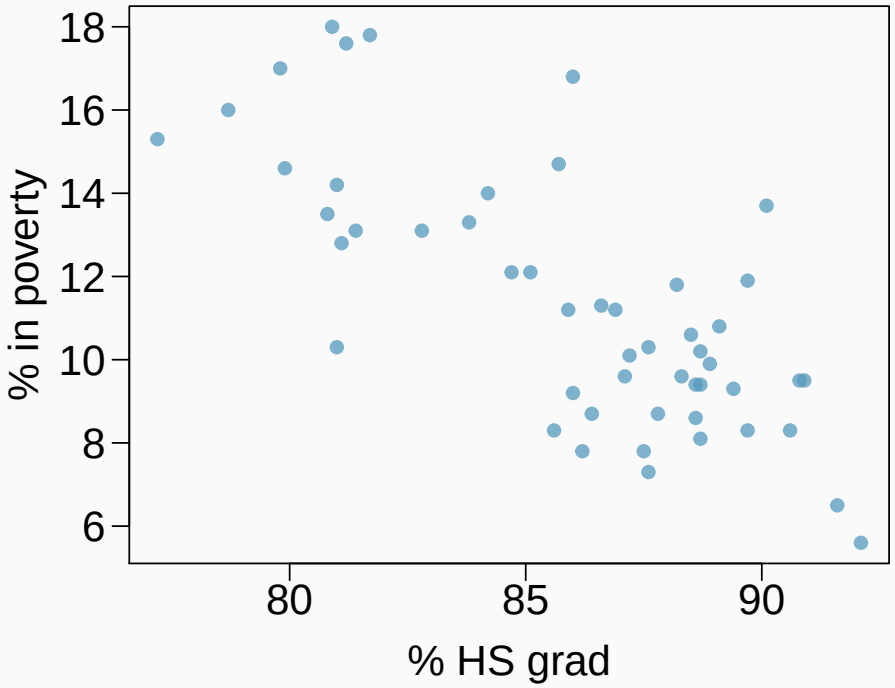
Response variable?

% in poverty

Explanatory variable?

Poverty vs. HS graduate rate

The *scatterplot* below shows the relationship between HS graduate rate in all 50 US states and DC and the % of residents who live below the poverty line (income below $23,050 for a family of 4 in 2012).



Response variable?

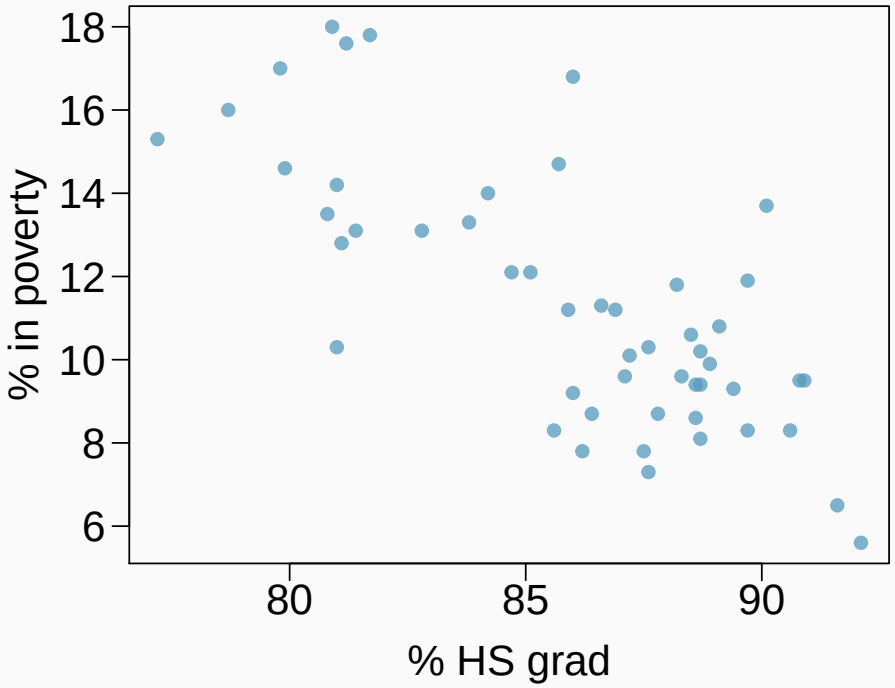
% in poverty

Explanatory variable?

% HS grad

Poverty vs. HS graduate rate

The *scatterplot* below shows the relationship between HS graduate rate in all 50 US states and DC and the % of residents who live below the poverty line (income below $23,050 for a family of 4 in 2012).



Response variable?

% in poverty

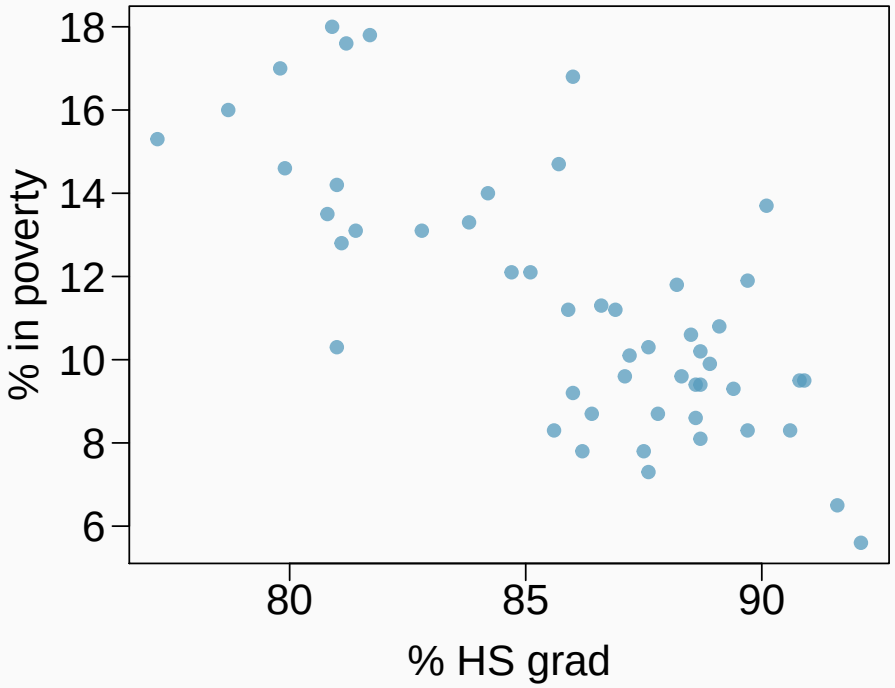
Explanatory variable?

% HS grad

Relationship?

Poverty vs. HS graduate rate

The *scatterplot* below shows the relationship between HS graduate rate in all 50 US states and DC and the % of residents who live below the poverty line (income below $23,050 for a family of 4 in 2012).



Response variable?

% in poverty

Explanatory variable?

% HS grad

Relationship?

linear, negative, moderately strong

Quantifying the relationship

- *Correlation* describes the strength of the *linear* association between two variables.

Quantifying the relationship

- *Correlation* describes the strength of the *linear* association between two variables.
- It takes values between -1 (perfect negative) and +1 (perfect positive).

Quantifying the relationship

- *Correlation* describes the strength of the *linear* association between two variables.
- It takes values between -1 (perfect negative) and +1 (perfect positive).
- A value of 0 indicates no linear association.

Guessing the correlation

Which of the following is the best guess for the correlation between % in poverty and % HS grad?

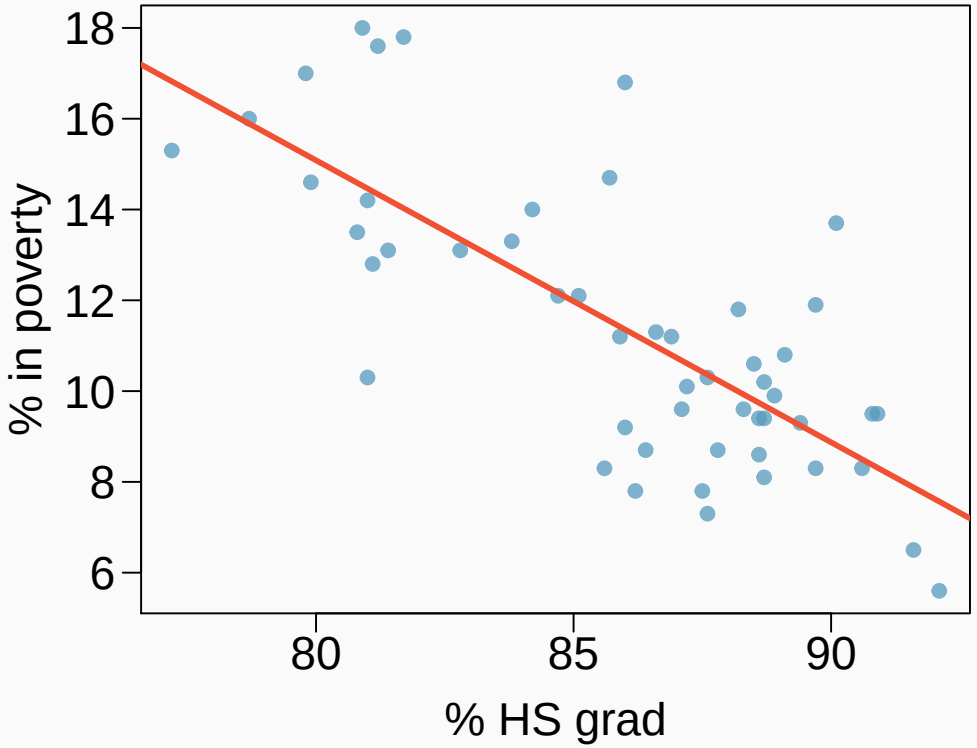
(a) 0.6

(b) -0.75

(c) -0.1

(d) 0.02

(e) -1.5



Guessing the correlation

Which of the following is the best guess for the correlation between % in poverty and % HS grad?

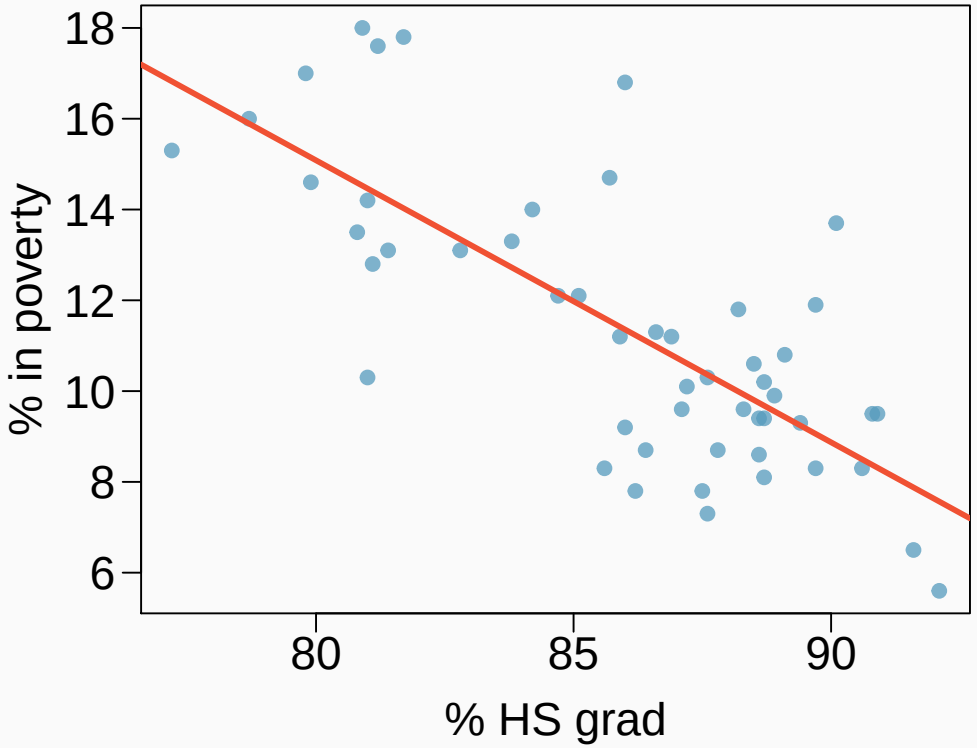
(a) 0.6

(b) *-0.75*

(c) -0.1

(d) 0.02

(e) -1.5



Guessing the correlation

Which of the following is the best guess for the correlation between % in poverty and % HS grad?

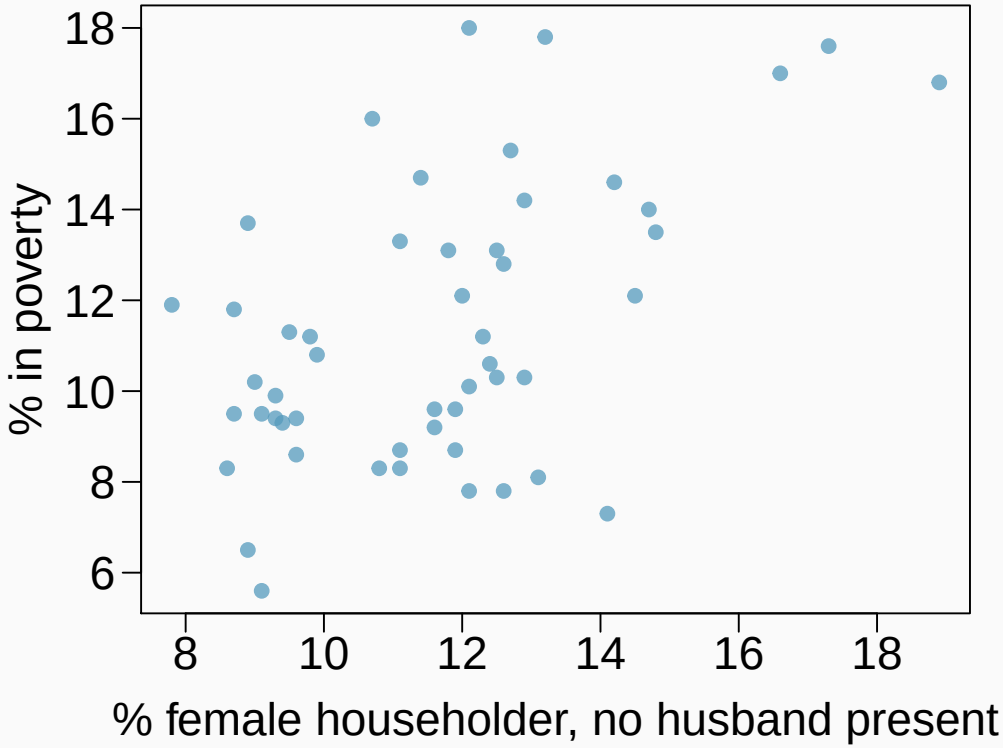
(a) 0.1

(b) -0.6

(c) -0.4

(d) 0.9

(e) 0.5



Guessing the correlation

Which of the following is the best guess for the correlation between % in poverty and % HS grad?

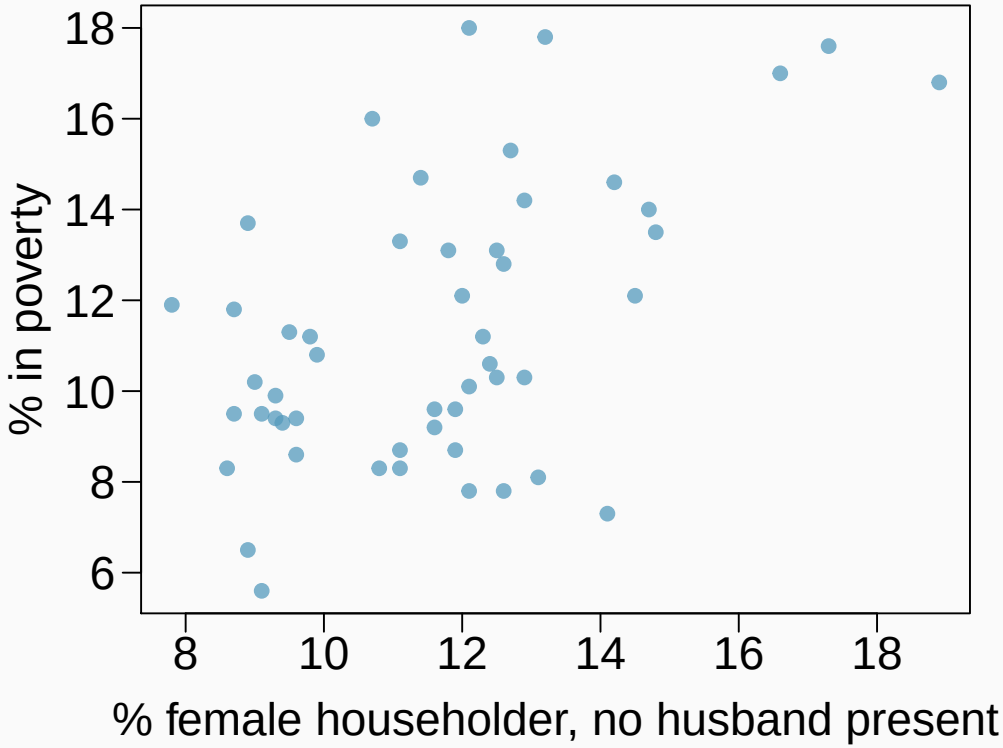
(a) 0.1

(b) -0.6

(c) -0.4

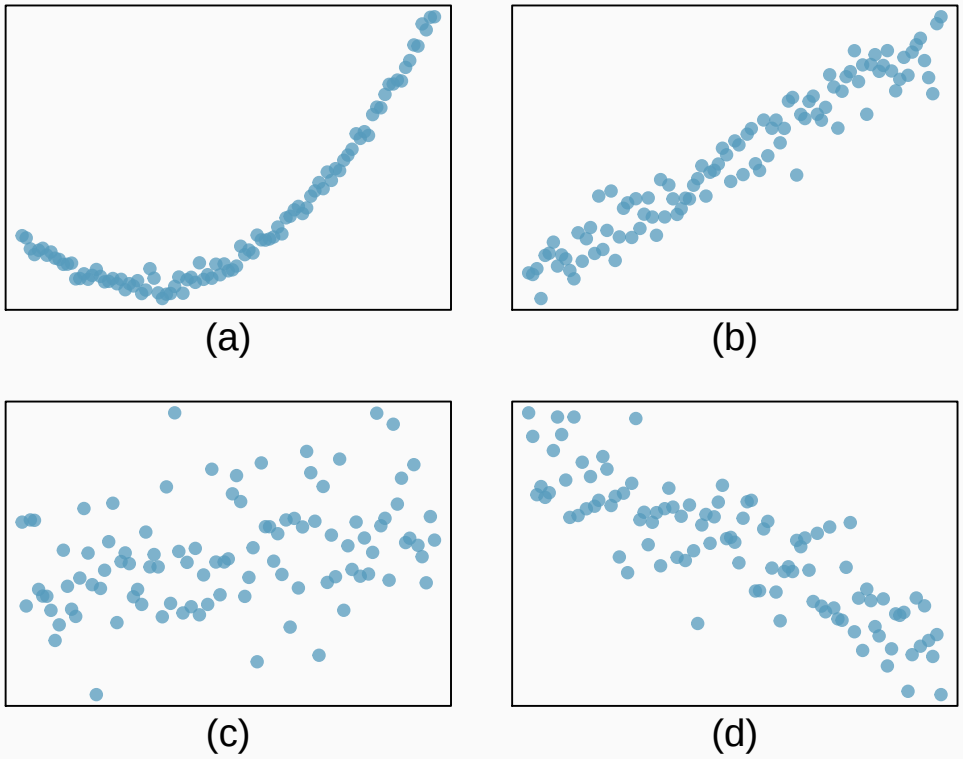
(d) 0.9

(e) *0.5*



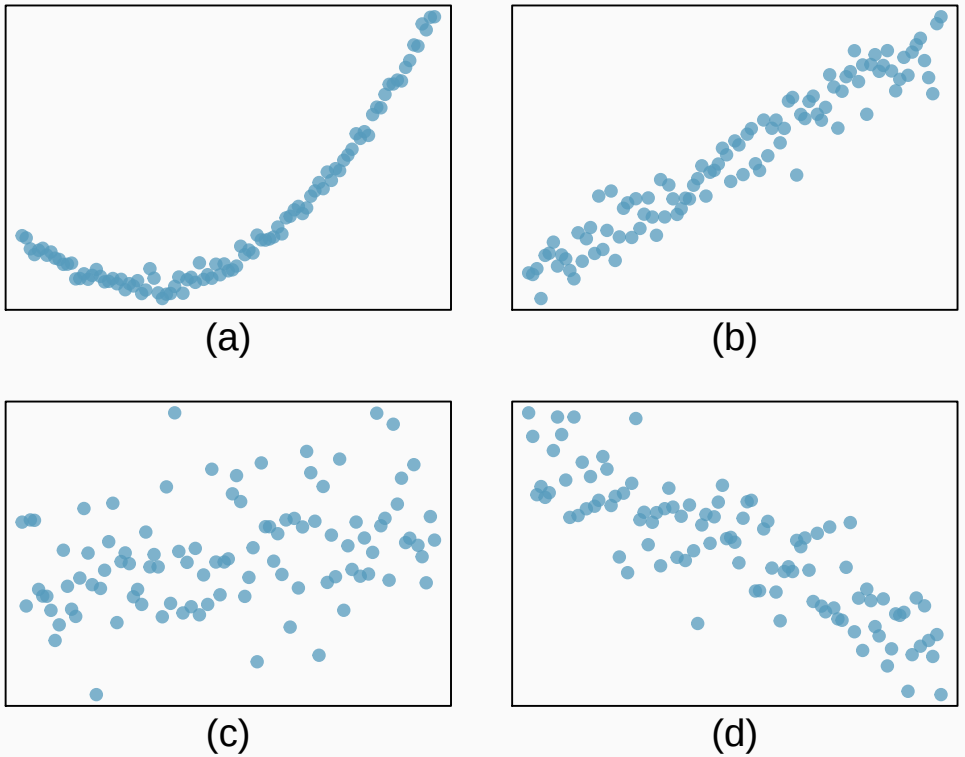
Assessing the correlation

Which of the following is has the strongest correlation, i.e. correlation coefficient closest to +1 or -1?



Assessing the correlation

Which of the following is has the strongest correlation, i.e. correlation coefficient closest to +1 or -1?



(b) → correlation means linear association

Credits

Content in **Confidence intervals** section and the slides with blue headers adapted from the chapter 4 and chapter 7 OpenIntro Statistics slides developed by Mine Çetinkaya-Rundel and made available under the CC BY-SA 3.0 license.