Linear regression, Sampling and Bias

Overview

Review of correlation and

Continuation of linear regression

Sampling and sampling bias

If there is time: Sampling distributions

Announcement

Homework 3 has been posted!

It is due on Gradescope on Sunday February 8th at 11pm

- **Be sure to mark each question on Gradescope!**

Jessica is going to have an R review session

- Time: Sunday 1-2pm
- Location: Bass L01A

Also, keep attending the practice sessions for more practice!

Announcement: mini-exam

The mini-exam will **in class** **on Thursday February 5th**

- Last 30 minutes of class

You should know/understand:

1. All the symbols we have used in class represent

   - E.g., A professor believes the average height of Yale students is 67 inches. How can you write this using the symbols we discussed in class?

     A: $\mu = 67$

2. Should be able to answer simple questions about the R code we have used

   - E.g., What does this R code do?
     boxplot(my_vec)

	Statistic	Parameter
Mean	$\bar{x}$	μ
Standard deviation	s	σ
Proportion	$\hat{p}$	π
Correlation	r	ρ
regression slope	b	β

Accommodations for the mini-exam

If you have accommodations for exams you have two options:

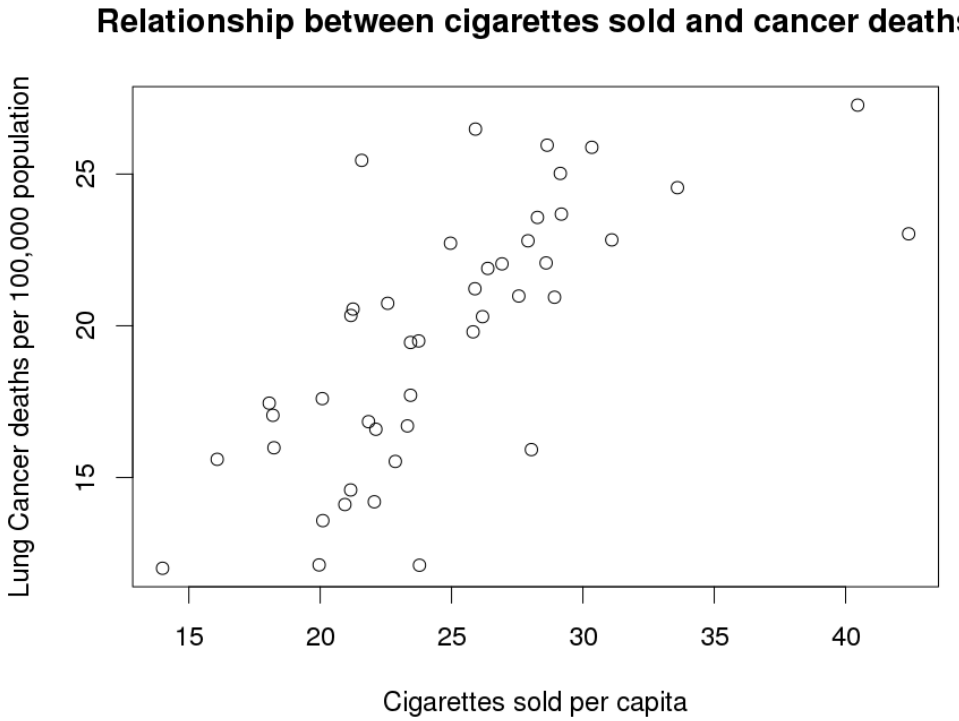
1. You can take the exam in class with everyone, and stay after the end of the class for additional time

- If you choose this option, please sit near the front of the class and let me know you are doing this so I know you can continue to work on the exam after the end of the class
  - We might also have to move if someone else is using the classroom after our class

2. You can take the exam with SAS. If you choose this option, please **contact SAS immediately** to schedule it with them, and let me know once you have done so

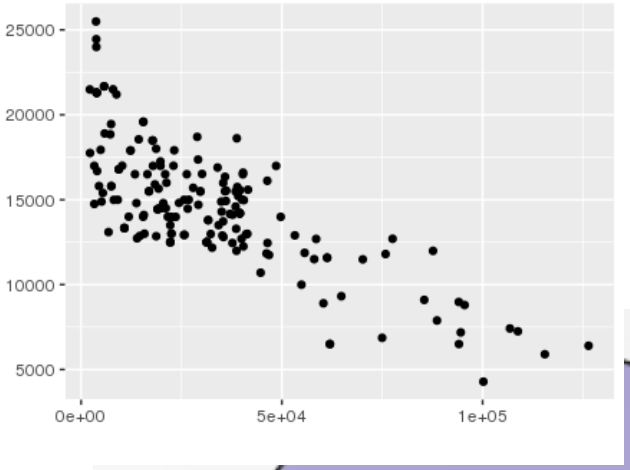
Review of correlation and linear regression

Review: scatter plots and the correlation coefficient

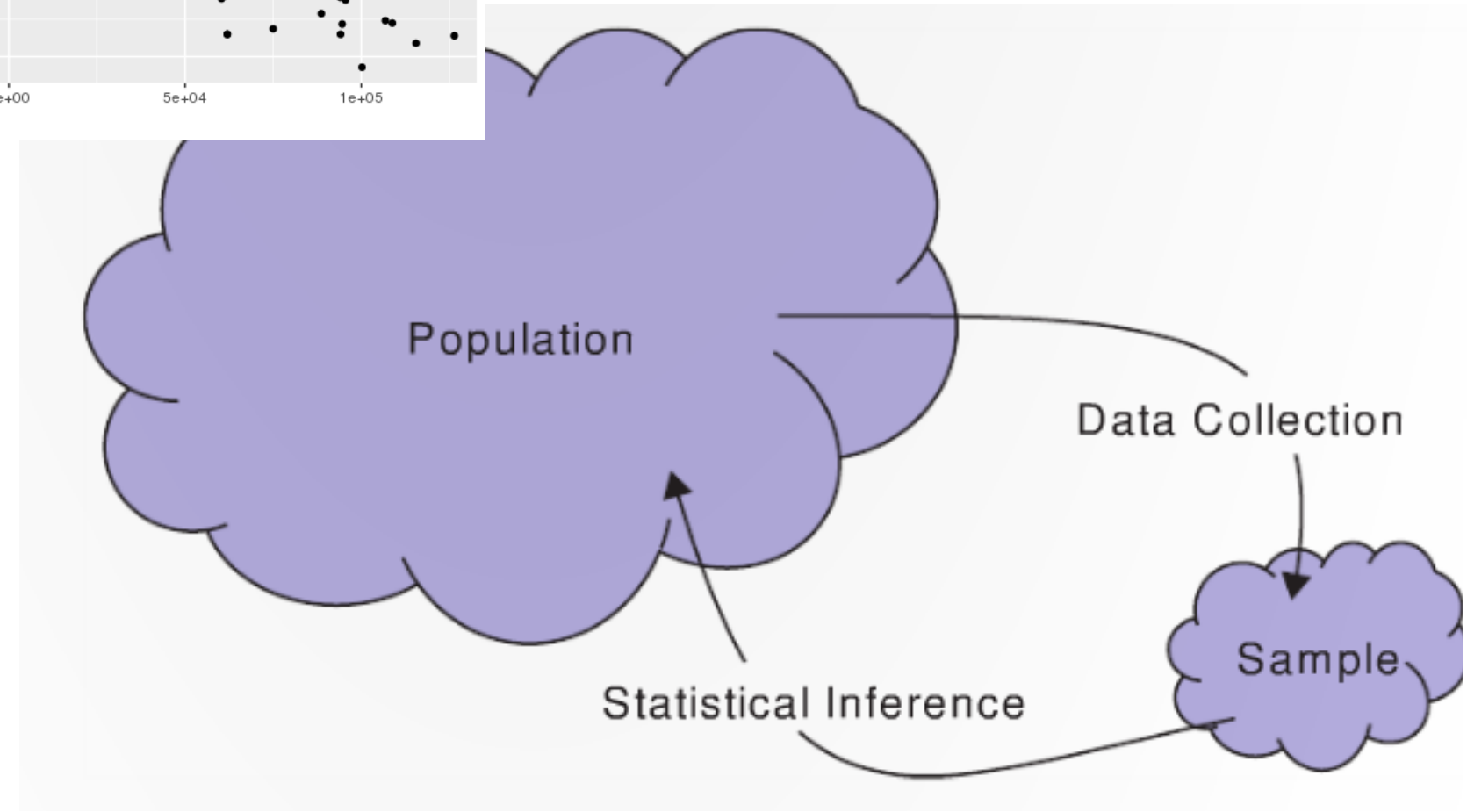


$$r = \frac{1}{(n-1)} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s_x} \right) \left(\frac{y_i - \overline{y}}{s_y} \right)$$

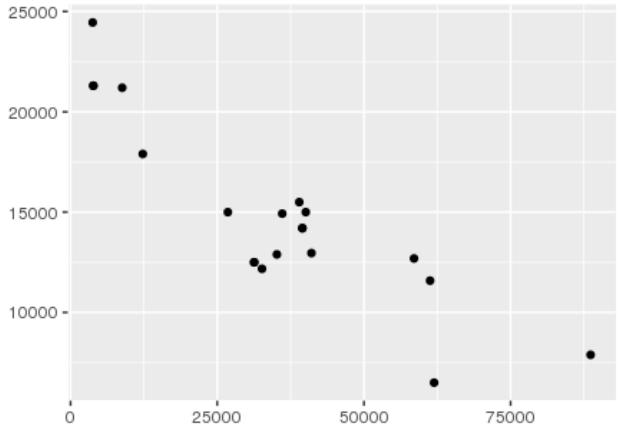
The **correlation** is measure of the strength and direction of a linear association between two variables



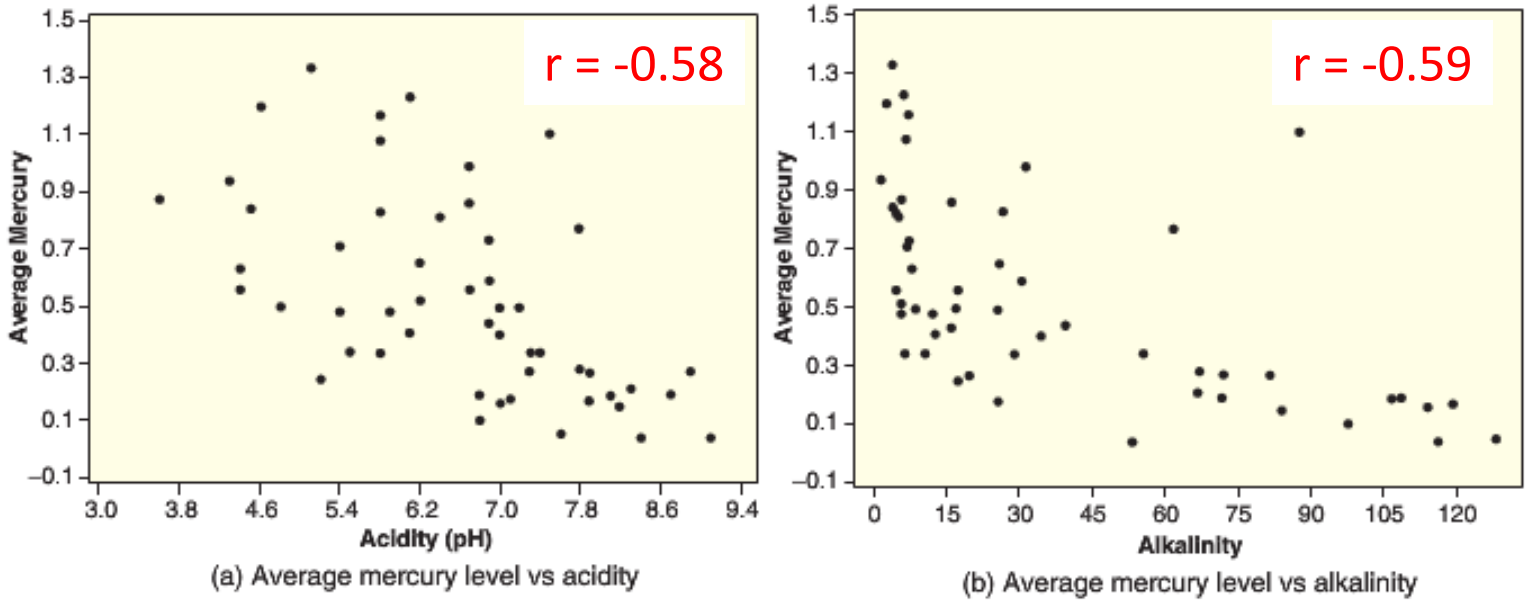
ρ parameter



r statistic

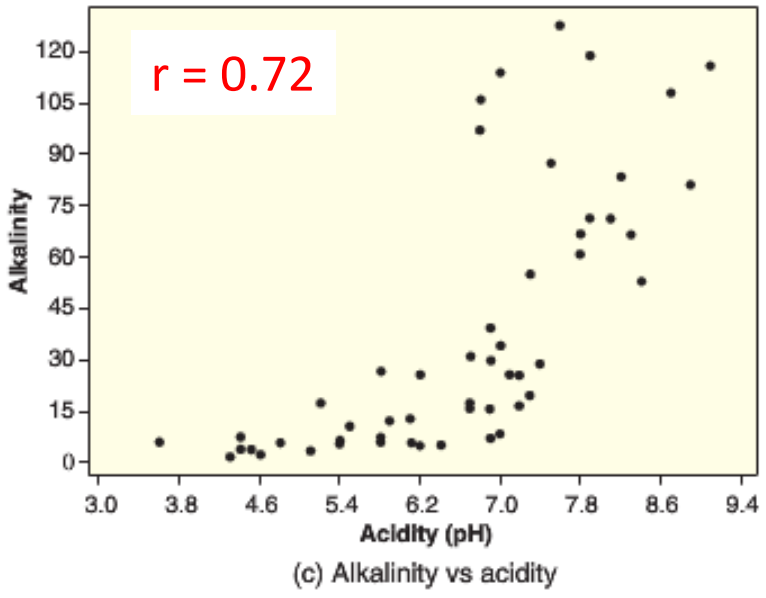


Florida lakes

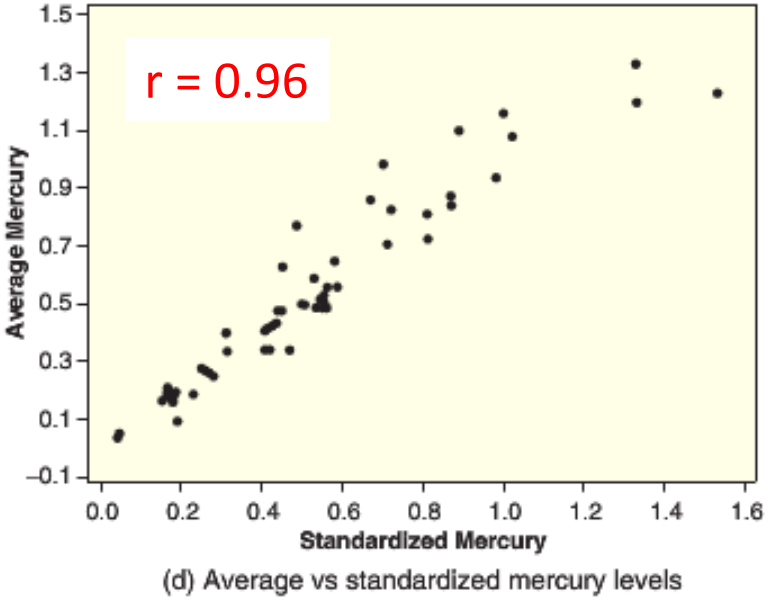


(a) Average mercury level vs acidity

(b) Average mercury level vs alkalinity



(c) Alkalinity vs acidity

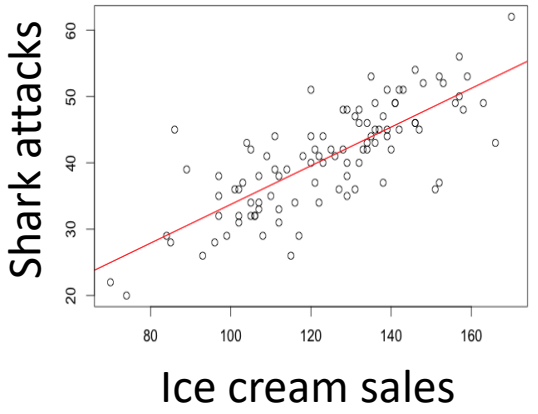


(d) Average vs standardized mercury levels

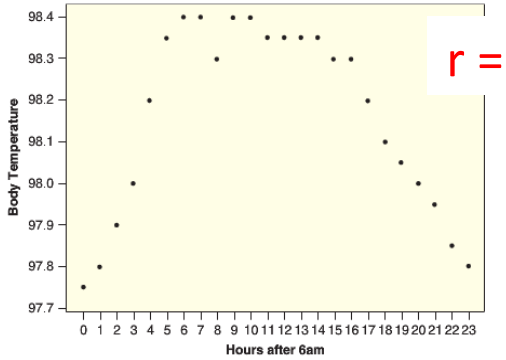
```
# create a scatter plot
plot(x, y)

# calculate the correlation
cor(x, y)
```

Correlation cautions

1. A strong positive or negative correlation does not (necessarily) imply a cause and effect relationship between two variables

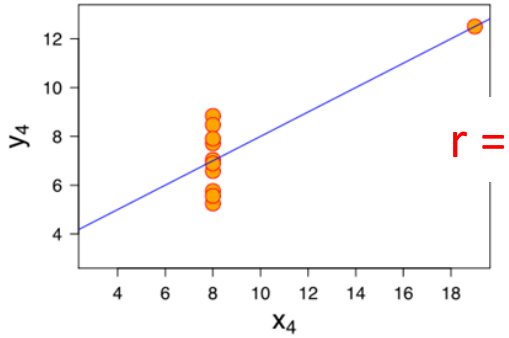


2. A correlation near zero does not (necessarily) mean that two variables are not associated. Correlation only measures the strength of a linear relationship



r = -0.08

3. Correlation can be heavily influenced by outliers. Always plot your data!



r = 0.81

Linear regression

Review: Regression

Regression is method of using one variable ***x*** to predict the value of a second variable ***y***

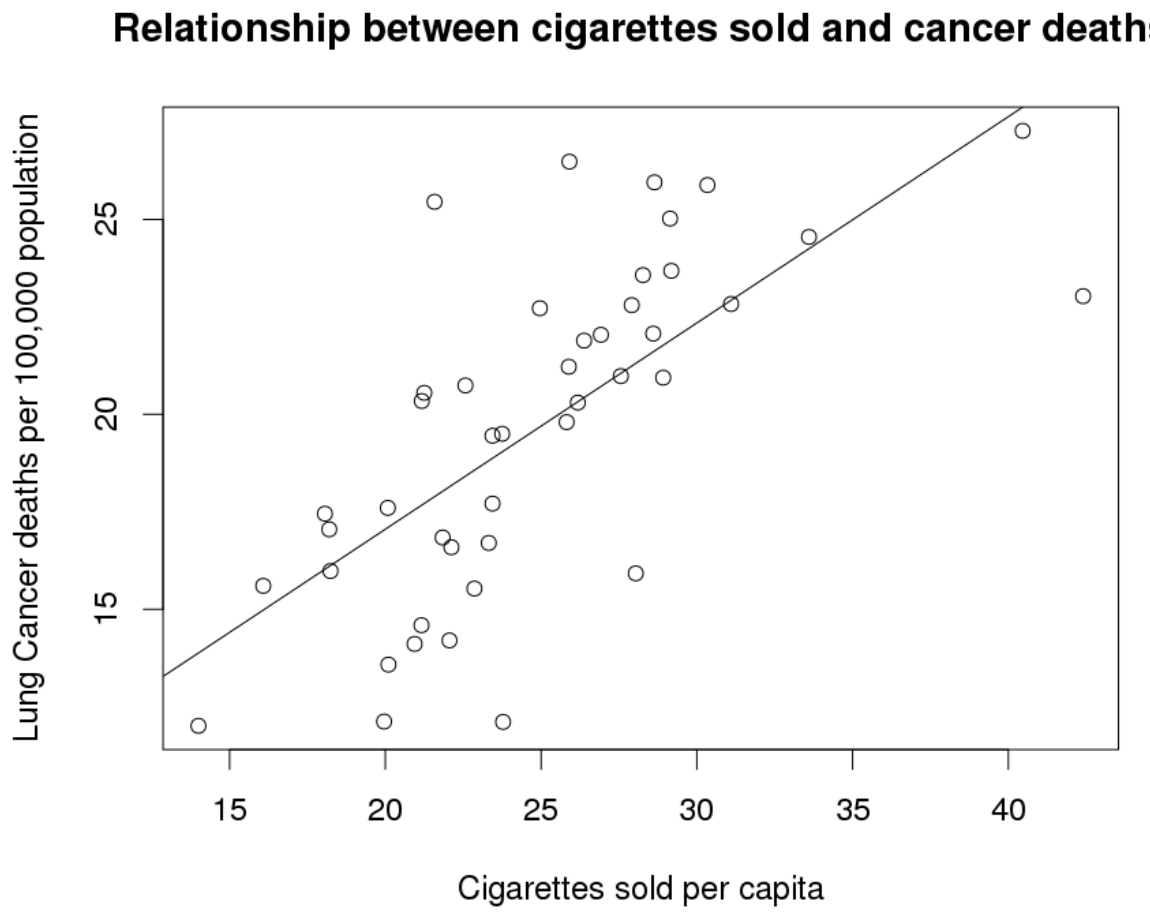
- i.e., $\hat{y} = f(x)$

In **linear regression** we fit a line to the data, called the **regression line**

$$\hat{y} = a + b \cdot x$$

$$Response = a + b \cdot Explanatory$$

Review: Cancer smoking regression line



$$\hat{y} = a + b \cdot x$$

R: `my_fit <- lm(y ~ x)`

`coef(my_fit)`

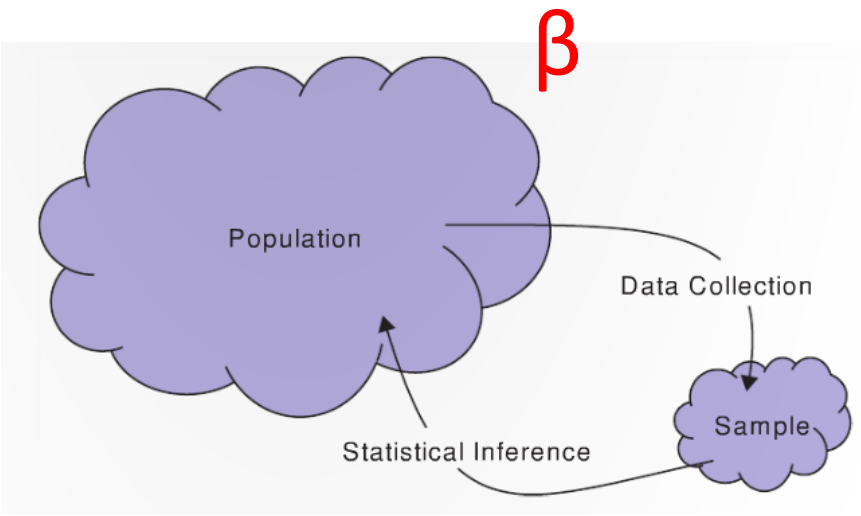
a = 6.47 b = 0.0053

$$\hat{y} = 6.47 + .0053 \cdot x$$

Notation

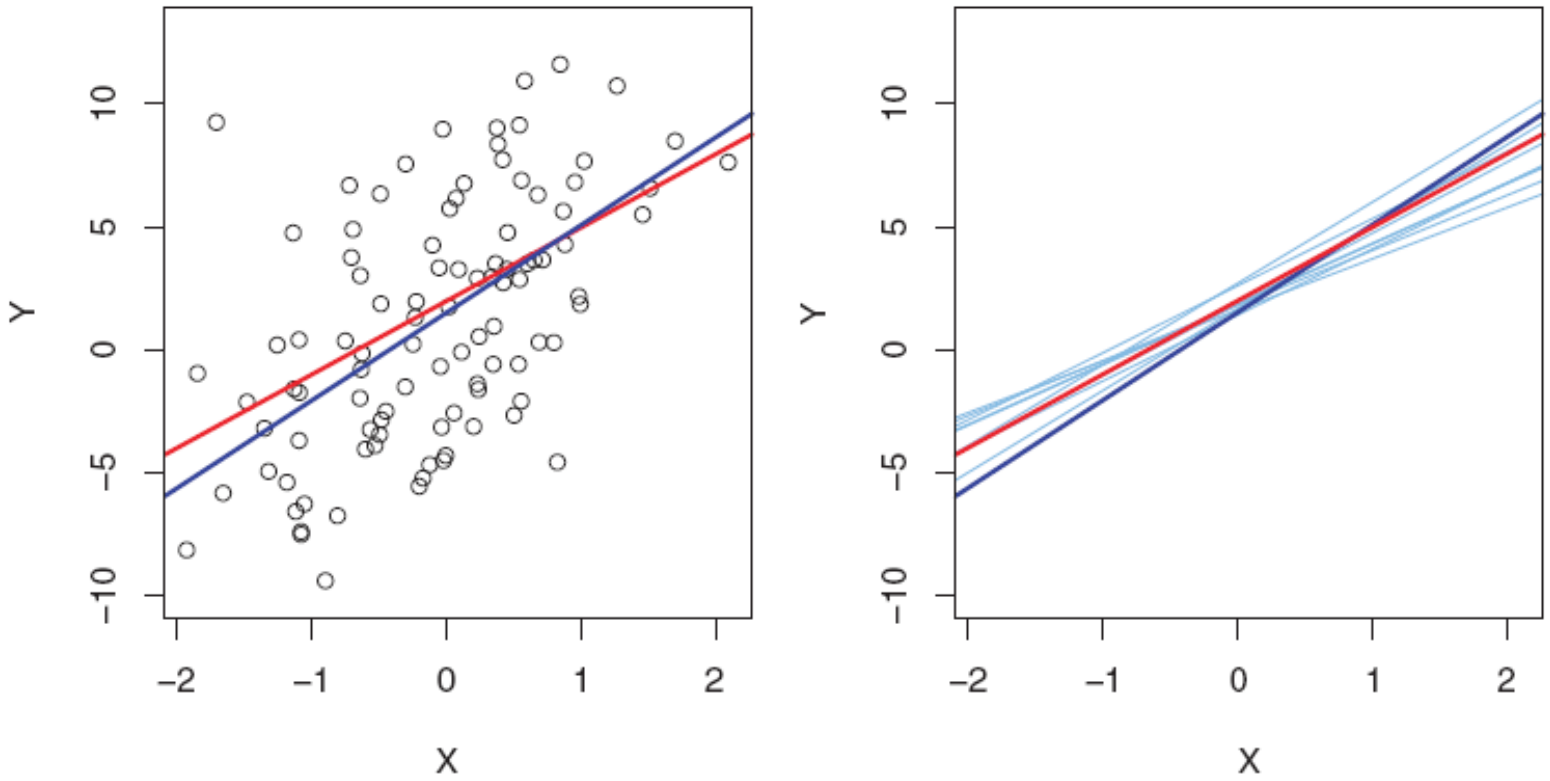
The letter **b** is typically used to denote the slope of the sample

The Greek letter **β** is used to denote the slope of the population



Population: β

Sample estimates: b

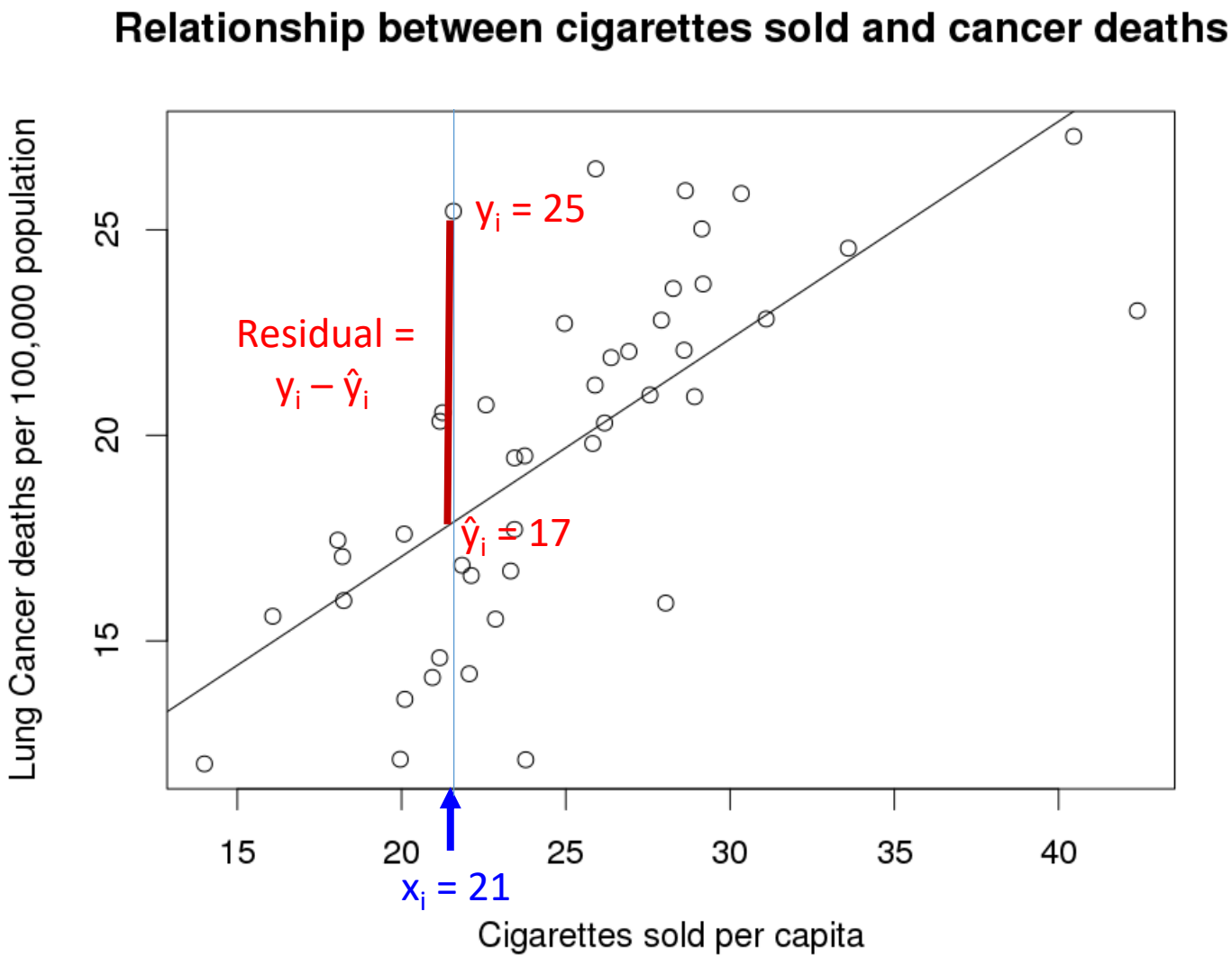


Residuals

The **residual** is the difference between an observed (y_i) and a predicted value ($\hat{y}_i$) of the response variable

$$Residual_i = Observed_i - Predicted_i = y_i - \hat{y}_i$$

Cancer smoking residuals



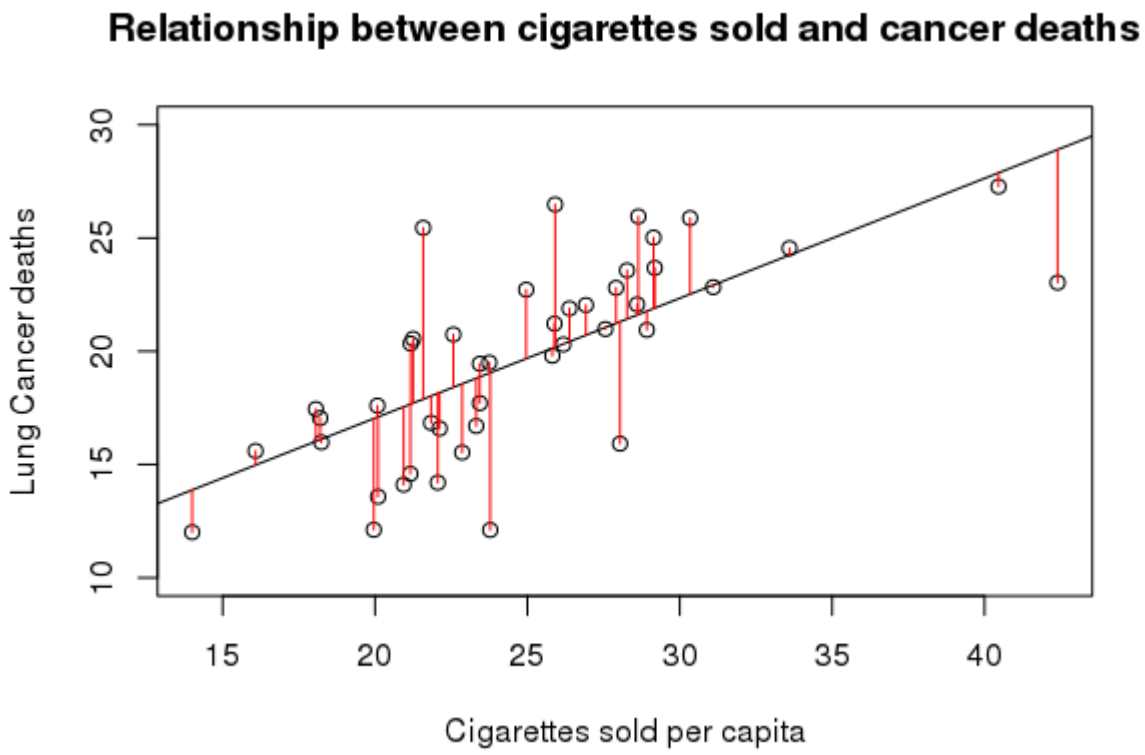
Cancer smoking residuals

$\hat{y} = 6.47 + 0.0053 \cdot x$

Cig per Capita (x)	Cancer obs (y)	Cancer pred ($\hat{y}$)	Residuals ($y - \hat{y}$)
1,820	17.05	16.10	0.95
2,582	19.80	20.13	-0.33
1,824	15.98	16.12	-0.14
2,860	22.07	21.60	0.47
3,110	22.83	22.93	-0.10
3,360	24.55	24.25	0.30
40,460	27.27	27.88	-0.61

Least squares line

The **least squares line**, also called '**the line of best fit'**, is the line which minimizes the sum of squared residuals



Find the line of best fit

Cancer smoking residuals

Cancer obs (y)	Cancer pred (ŷ)	Residuals (y - ŷ)	Residuals² (y - ŷ)²
17.05	16.10	0.95	0.90
19.80	20.13	-0.33	0.11
15.98	16.12	-0.14	0.02
22.07	21.60	0.47	0.22
22.83	22.93	-0.10	0.01
24.55	24.25	0.30	0.09
27.27	27.88	-0.61	0.37
23.57	21.24	2.14	4.59

$$\hat{y} = a + b \cdot x$$

Find the a and b

That minimizes the sum of the squared residuals

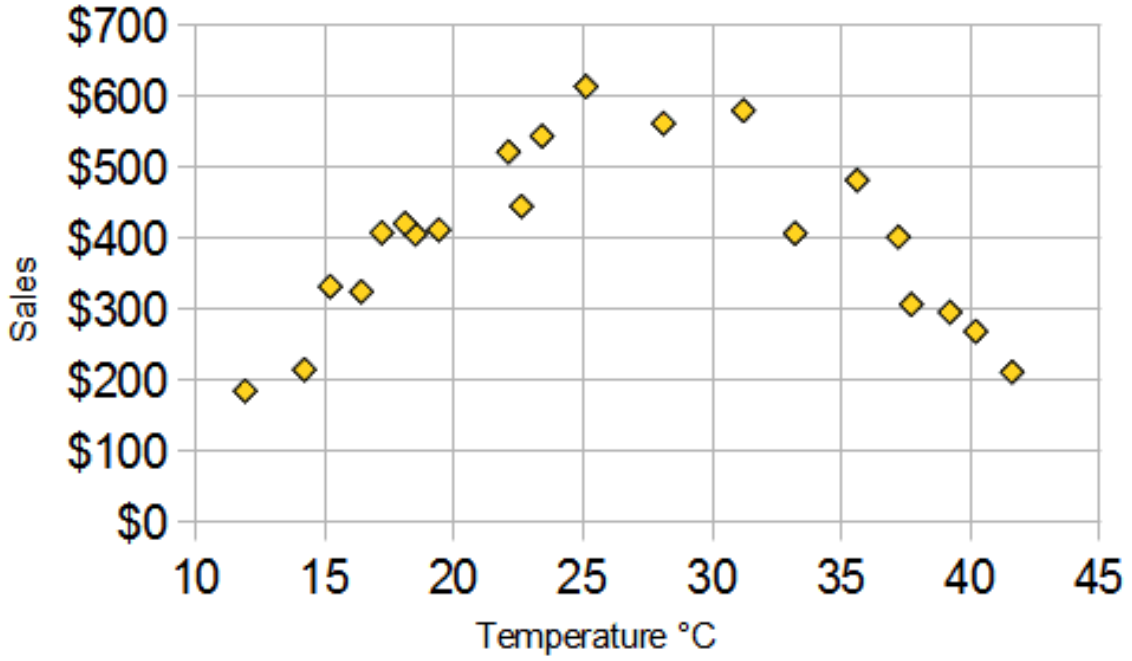
Regression caution # 1

Avoid trying to apply the regression line to predict values far from those that were used to create the line

- i.e., do not extrapolate too far

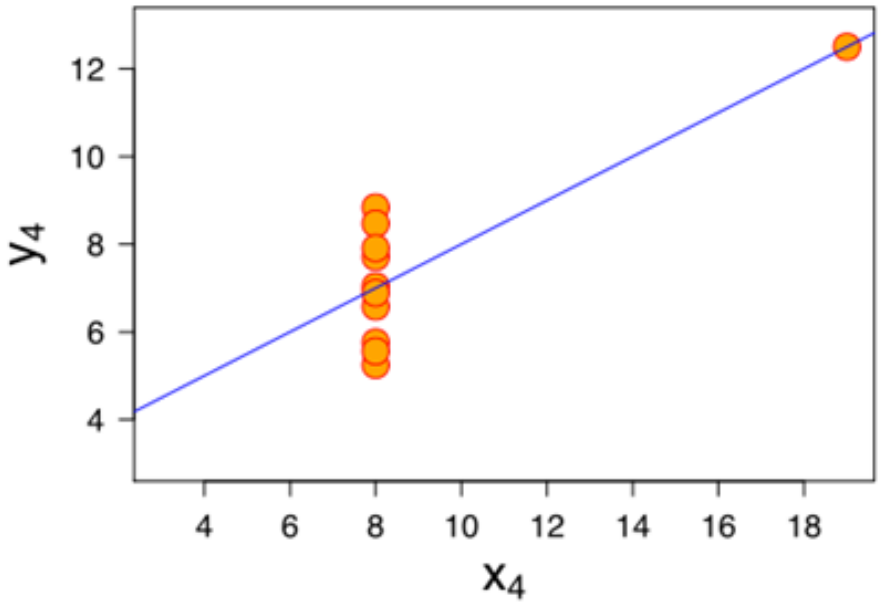
Regression caution # 2

Plot the data! Regression lines are only appropriate when there is a linear trend in the data



Regression caution #3

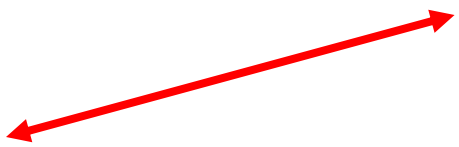
Be aware of outliers – they can have an huge effect on the regression line



Linear regression in R

Regression lines in R

```
# load the data
   load("states_smoking.rda")

# create a scatter plot and calculate the correlation        Note:   plot(x, y)
   plot(smoking$CIG, smoking$LUNG)

# fit a regression model      Note:    lm(y ~ x)
   lm_fit <- lm(smoking$LUNG ~ smoking$CIG)

# get the a and b coefficients
  coef(lm_fit)

# add a regression line to the plot
  abline(lm_fit)
```



Concepts for the relationship between two quantitative variables

A **scatterplot** graphs the relationship between two variables

The **correlation** is measure of the strength and direction of a linear association between two variables

- Value between -1 and 1

In **linear regression** we fit a line to the data, called the **regression line**

- We get coefficients for the slope (b) and the y-intercept (a)

The **residual** is the difference between an observed (y_i) and a predicted value ($\hat{y}_i$) of the response variable

- The regression line minimizes the sum of squared residuals

Sampling

Where do samples/data come from?

Example: sampling 100 sprinkles

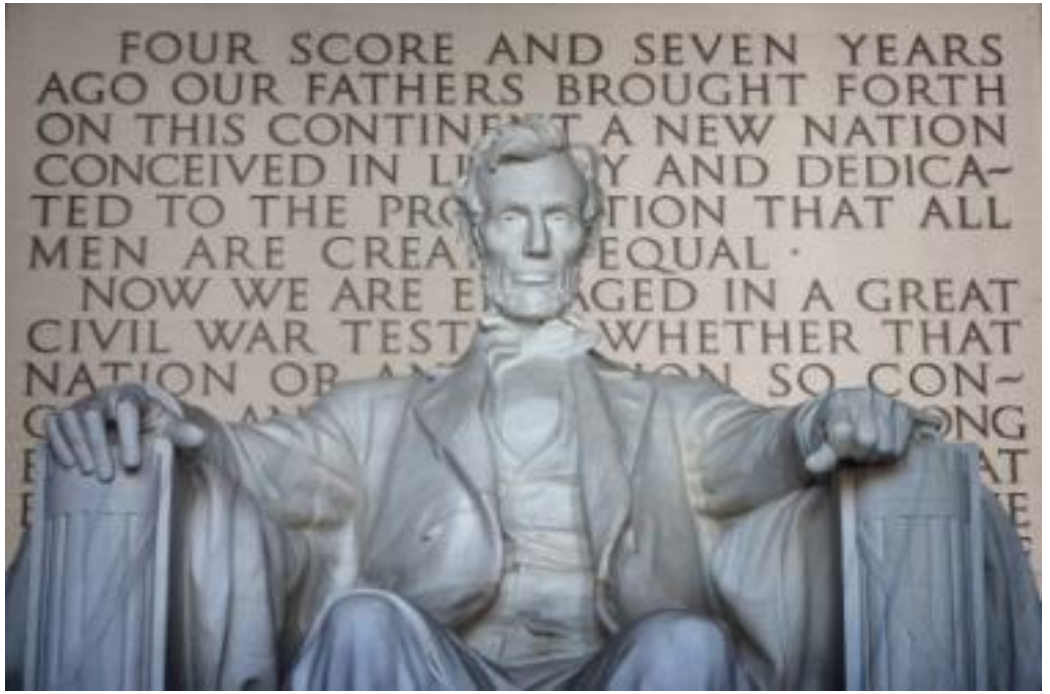
1	orange
2	red
3	green
4	white
5	white
6	white
7	white
8	white
9	red

The ***sample size*** (n) is the number of items in the sample
What is ***n*** in the sprinkle example?

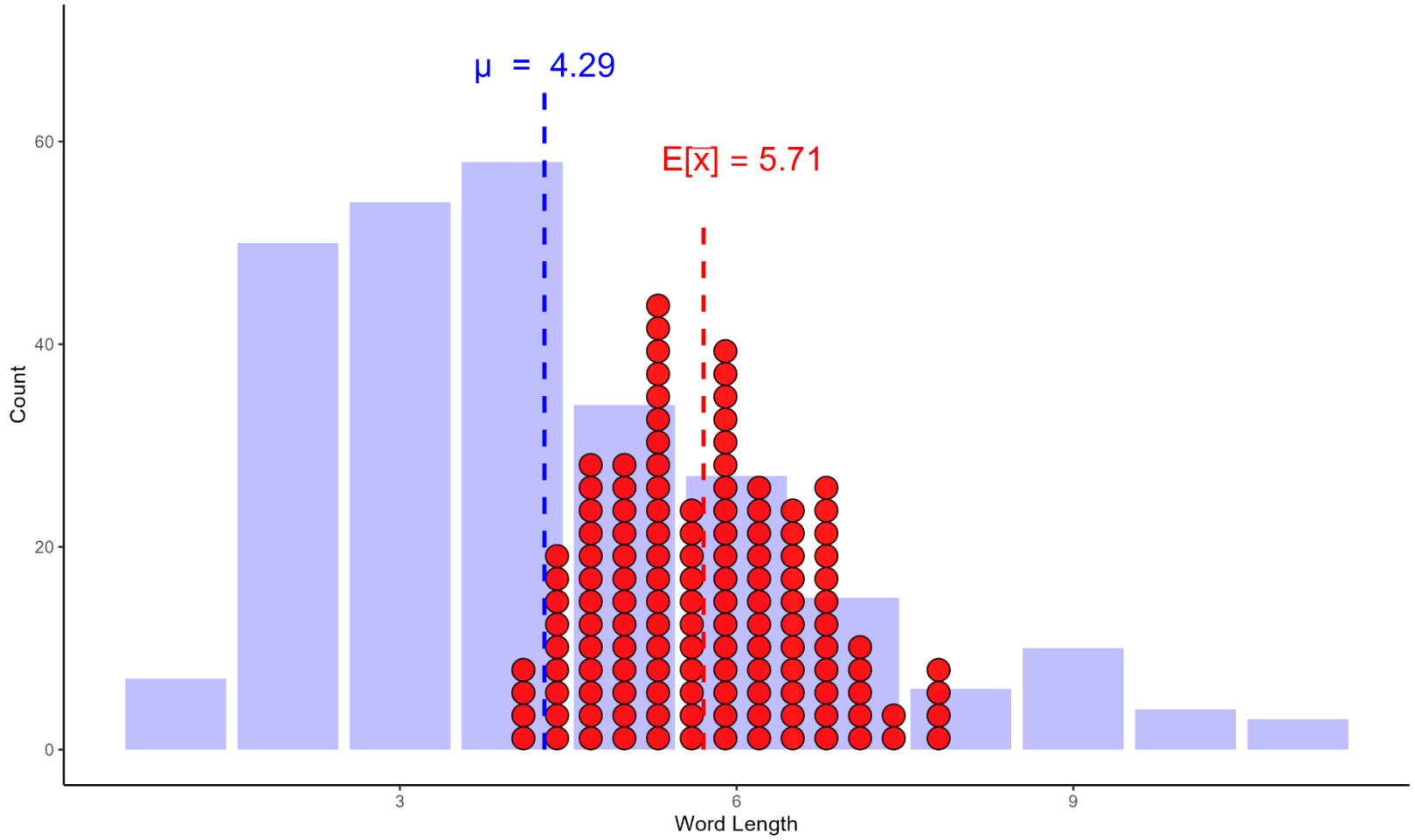
Sampling from the Gettysburg address

You filled out Gettysburg sampling worksheet survey on Canvas where you will randomly sample 10 words from the Gettysburg address

You reported the mean word $\bar{x}$ length of your 10 words



Bias and the Gettysburg address word length distribution

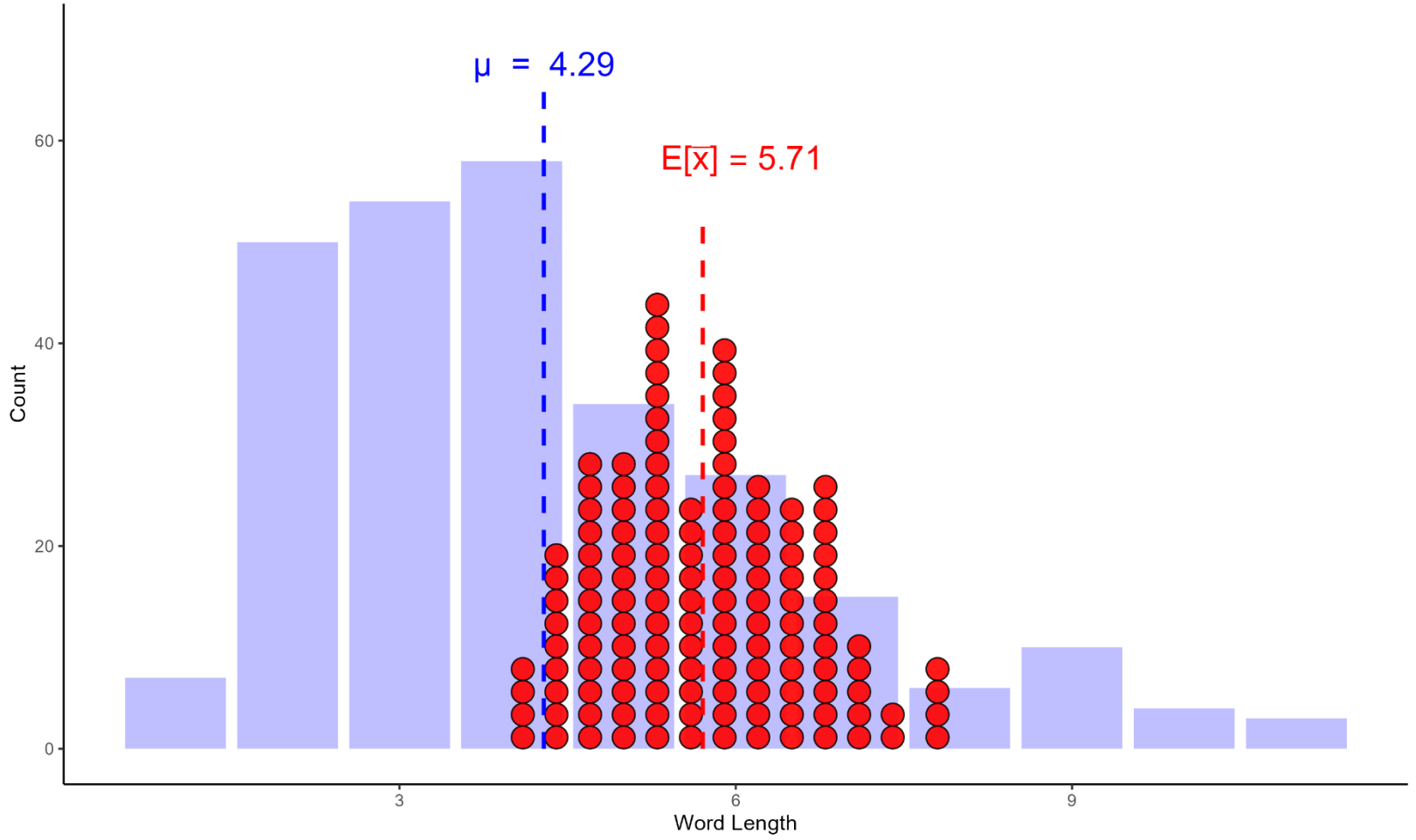


Bias and the Gettysburg address word length distribution

Bias is when the average statistic values does not equal the population parameter

Here:

$$E_s[\bar{x}] \neq \mu$$



Bias

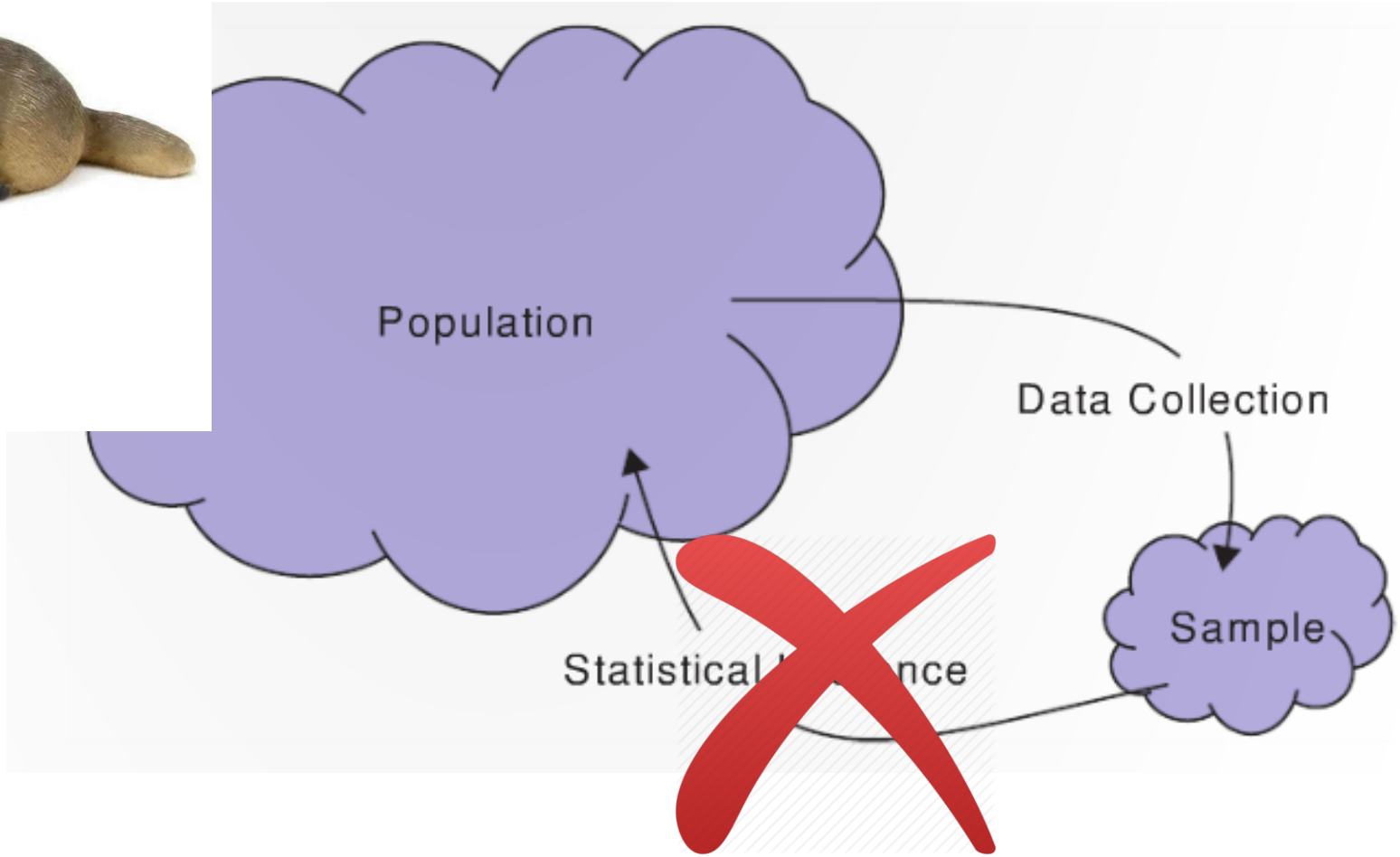
Sampling bias exists when the method of collecting the data causes the sample to inaccurately reflect the population

This leads to ***biased statistics*** where our average statistic value does not equal the parameter value

- E.g., $E_s[\bar{x}] \neq \mu$

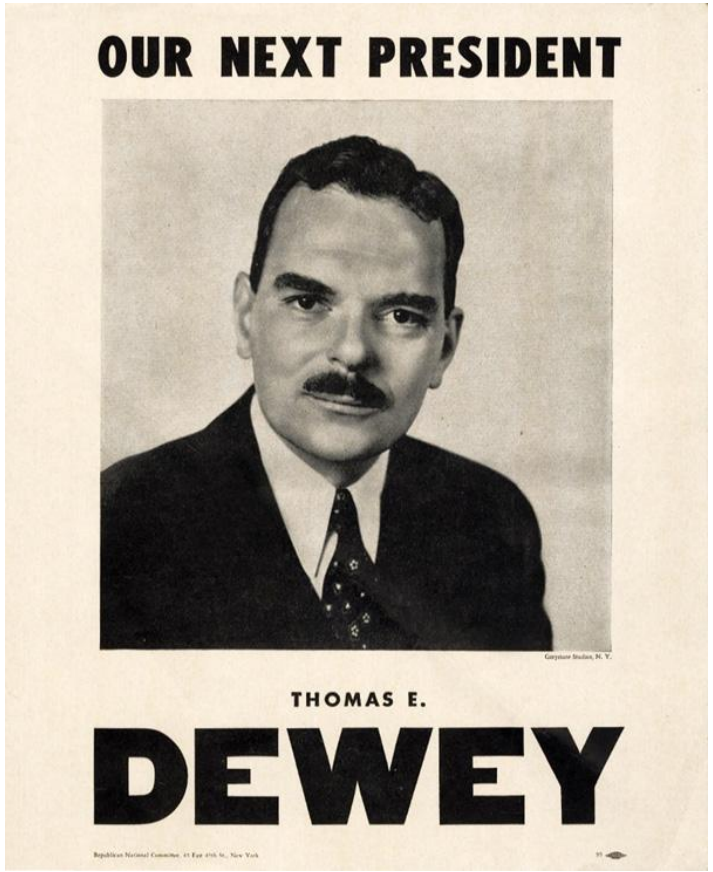
Statistical bias

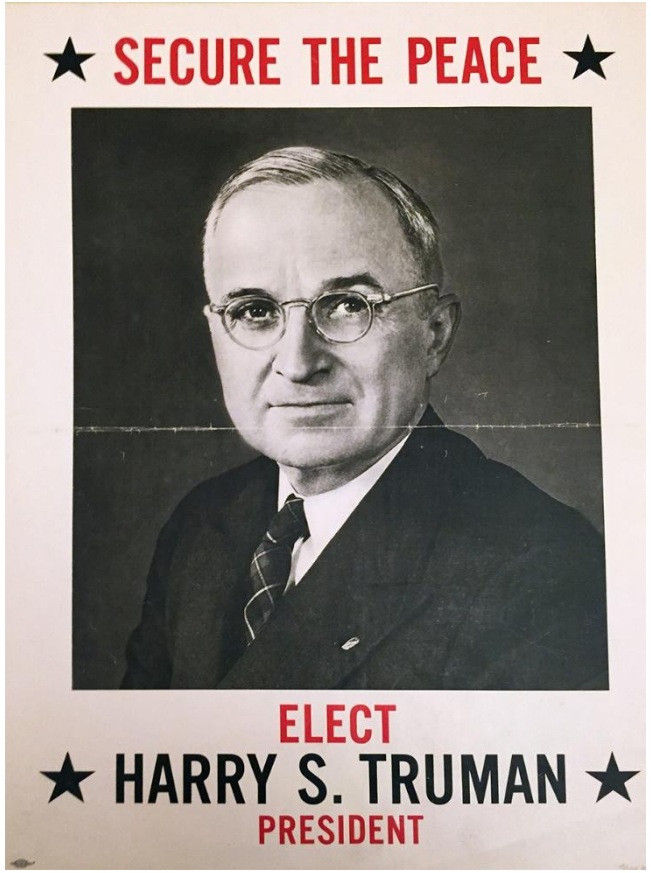
μ



$\bar{x}$

1948 US election





Newspaper title: Dewey Defeats Truman (1948)

The newspaper was published before the conclusion of the 1948 presidential election

The results were based on a large telephone poll which showed Dewey sweeping Truman

However, Harry S. Truman won the election

Q: What went wrong?

Basic questions for sampling

What is the population?

What is the sample?

Do they differ in a meaningful way?

To prevent bias: use simple random sample!

Simple random sample: each member in the population is equally likely to be in the sample

Allows for generalizations to the population!

Soup analogy

How do we select a random sample?

Mechanically:

- Flip coins
- Pull balls from well mixed bins
- Deal out shuffled cards, etc.

Use computer programs

Bias or no bias?

1948 US election: Dewey vs. Truman

Suppose there was a poll for the Truman/Dewey election that had randomly chosen 6,000 people from all voters in the USA and calculated who they voted for

In the spring 2013 Hampshire College launched a survey of alums

Via email, the College **invited 8,160 alums to fill out an online questionnaire**

A total of 1,920 surveys were completed, yielding a response rate of 24%

Alumni Survey Results

Hampshire College

As part of a strategic-planning process, in spring 2013 Hampshire College launched a survey of alums. Via email, the College invited 8,160 alums to fill out an online questionnaire administered by the campus's Alumni and Family Relations and Institutional Research offices. A total of 1,920 surveys were completed, yielding a response rate of 24%.

Note: The percentages in the data (below) are based on the number of responses received for each question.

To what extent do you agree with the following statements?

Strongly Agree or Agree

Hampshire encouraged me to think and work independently	Hampshire encouraged me to come up with innovative ideas and solutions	Hampshire improved my ability to synthesize information from across disciplines	Hampshire helped shape me into a life-long learner
99%	95%	96%	95%

Please rate your student experience at Hampshire.

65% of our alumni earn advanced degrees within ten years of graduating.

1 in 7 alumni holds a Ph.D. or other terminal degree.

Hampshire ranks in the **top 1%** of colleges nationwide in the % of grads that go on to earn doctorates.

26% of our graduates have started their own business or organization.

"

Hampshire does a great job fostering the ability to ask good questions and to look at ideas with a critical lens.

Hampshire has encouraged me to be more engaged, socially aware and more of a critical thinker than my peers.

I feel more able to adapt to a range of environments because Hampshire taught me skills and ideas rather than just knowledge.

"

Yelp reviews of restaurants?

An anonymous survey randomly select 6,000 people and asked them if have they used an elicit drug in the past month?

https://www.billoreilly.com/poll-center

The way you frame the question matters!

Quinnipiac University conducted two polls on November 5, 2015

First poll asked: do you support “stricter gun control laws”?

- Yes = 46% No = 51% Difference = -5%

Second poll asked: do you support “stricter gun laws”?

- Yes = 52% No = 45% Difference = 7%

How could this affect the newspaper headlines?

Also see textbook section 1.2:

- “If you had to do it over again, would you have children?”

Practicalities…

It might not be feasible to randomly select equally from all members of a population

This might not be a problem as long as the sample is representative of the population

Example: If we wanted to know proportion of left-handed people in the US, randomly sampling Yale students might be sufficient

Need to think carefully to avoid bias!

Statistics requires thought!

Use your own reasoning:

- What is the population I am interested in?
- Does the sample reflect the population of interest?
- Be your own worst critic!

Questions about statistical bias?

From now on we are going to assume no bias!

Our statistic values, on average, reflect the parameters