Regression continued, bias and sampling distributions

Overview

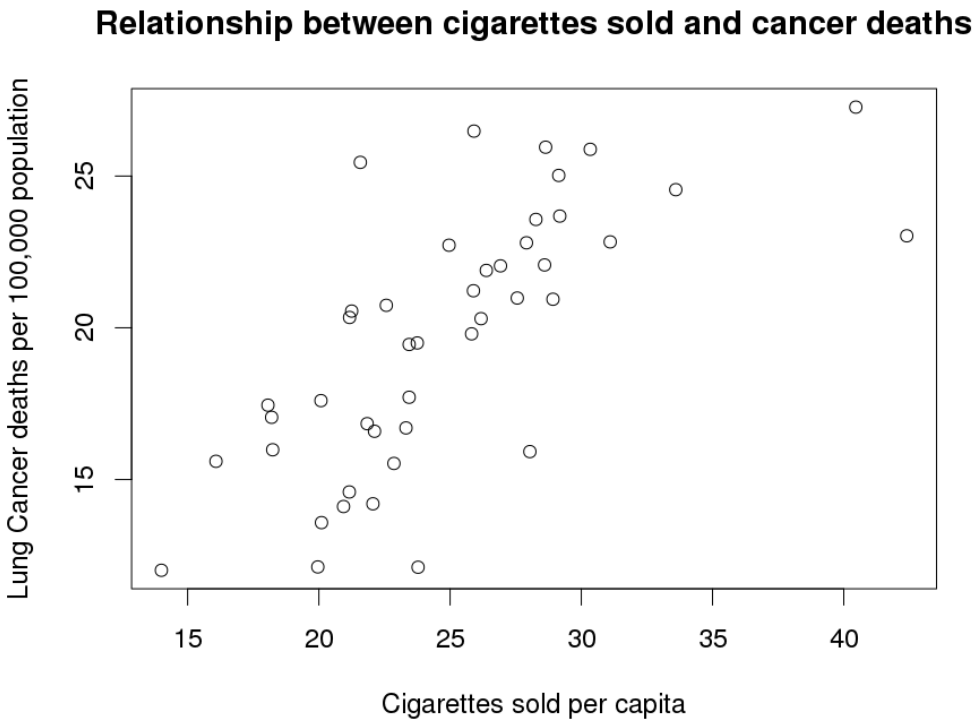
Questions about homework 2?

Review and continuation of simple linear regression

Sampling and bias

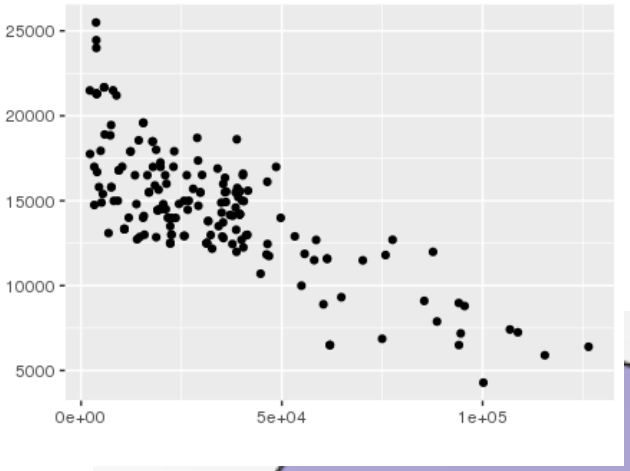
Sampling distributions

Review: scatter plots and the correlation coefficient

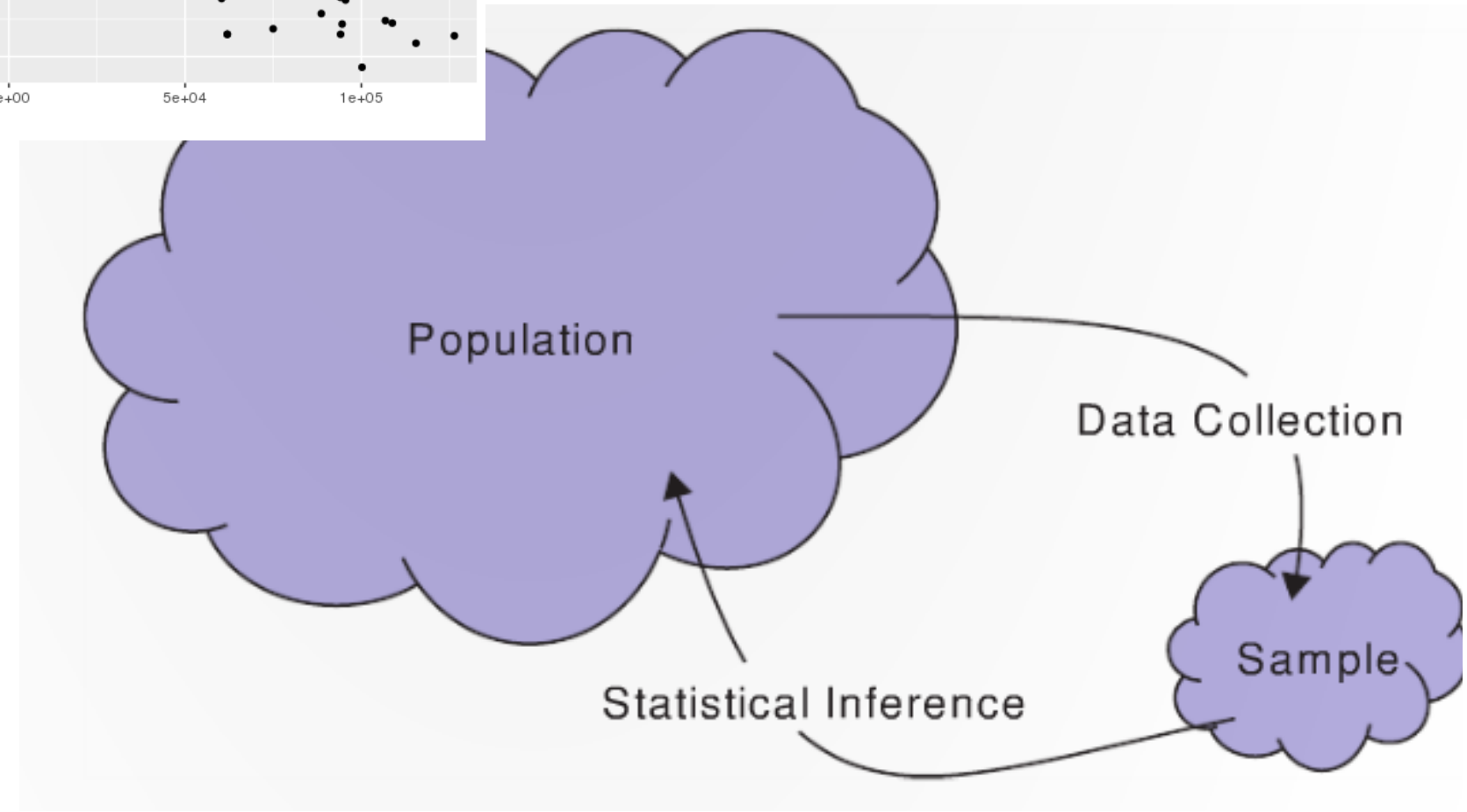


$$r = \frac{1}{(n-1)} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s_x} \right) \left(\frac{y_i - \overline{y}}{s_y} \right)$$

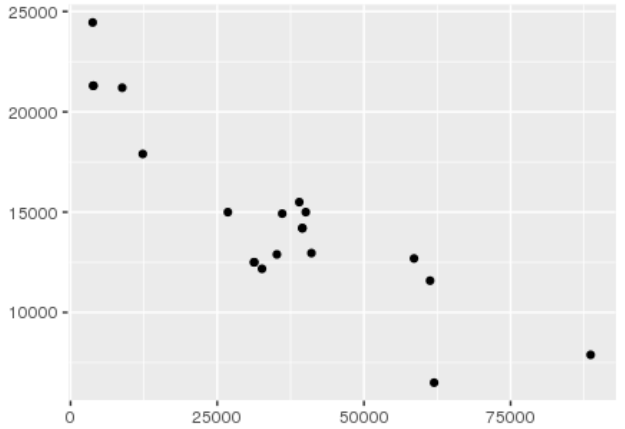
The **correlation** is measure of the strength and direction of a linear association between two variables



ρ parameter



r statistic



Regression

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

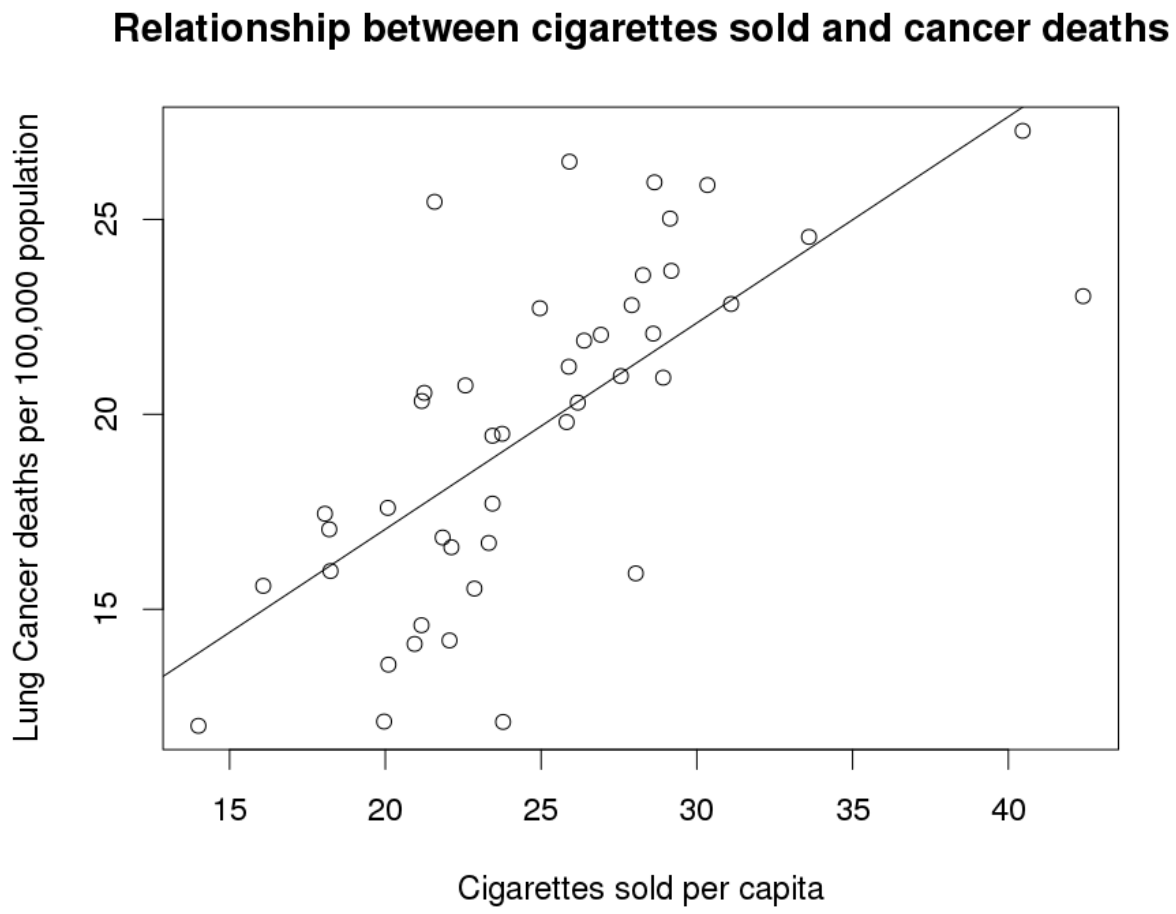
- i.e., $\hat{y} = f(x)$

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

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

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

Cancer smoking regression line



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

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

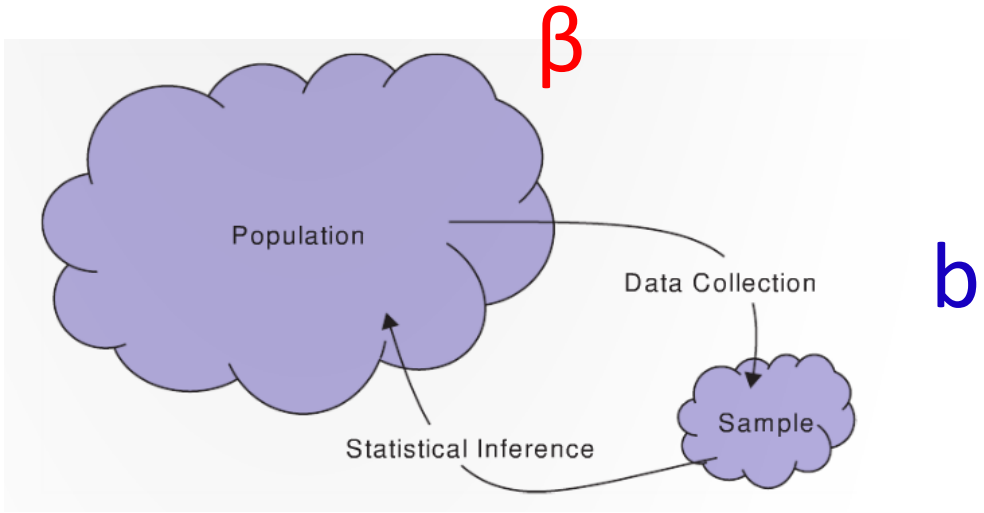
a = 6.47 b = 0.53

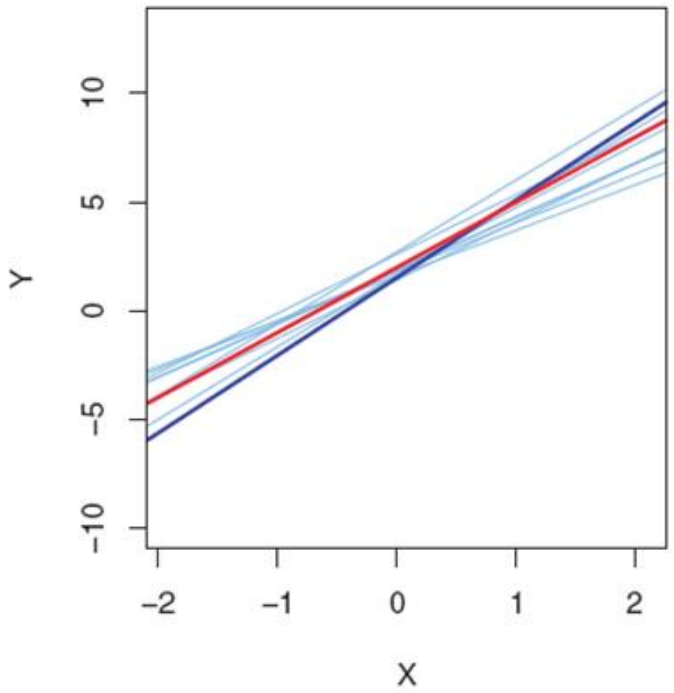
$$\hat{y} = 6.47 + .53 \cdot x$$

Notation

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

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



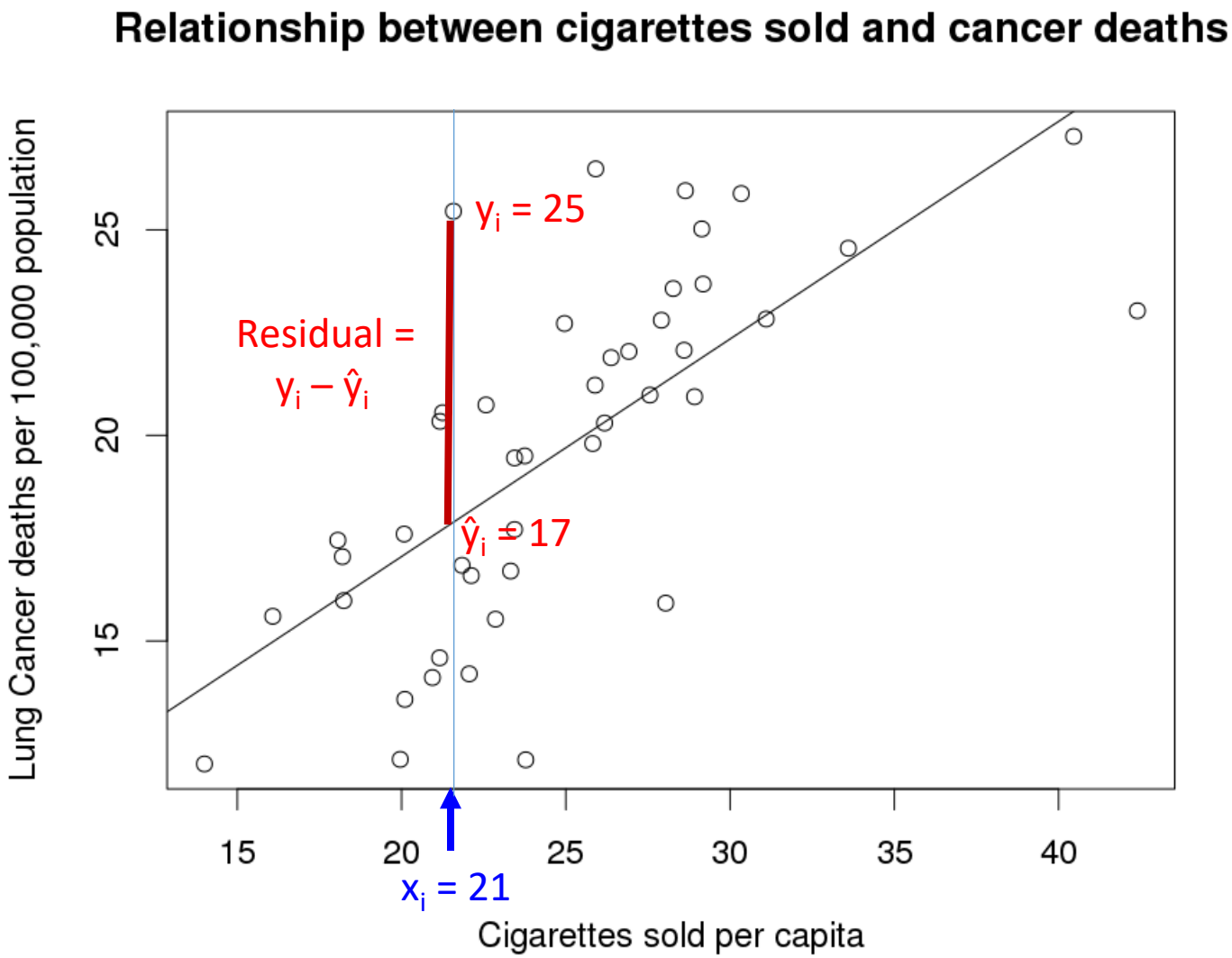


Residuals

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



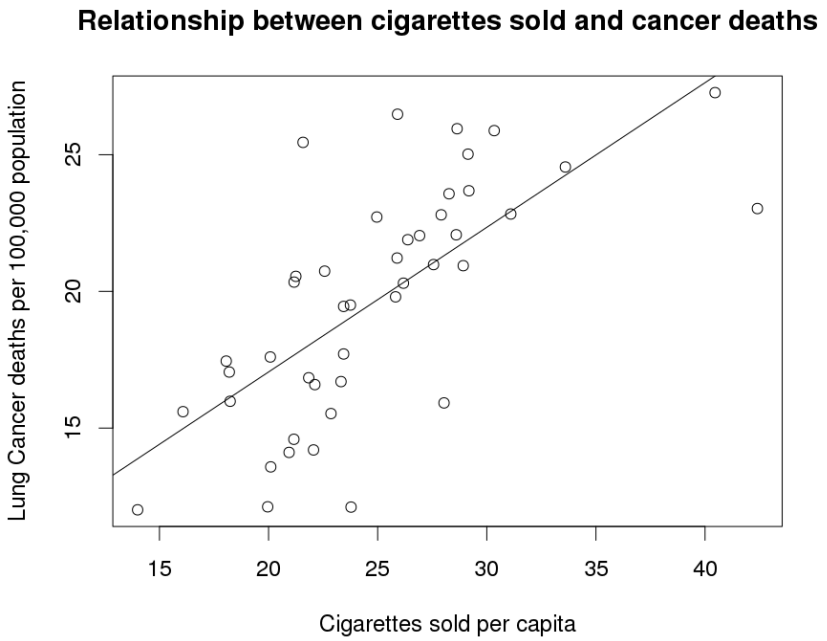
Cancer smoking residuals

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

Cancer obs (y)	Cancer pred (ŷ)	Residuals (y - ŷ)
17.05	16.10	0.95
19.80	20.13	-0.33
15.98	16.12	-0.14
22.07	21.60	0.47
22.83	22.93	-0.10
24.55	24.25	0.30
27.27	27.88	-0.61
23.57	21.24	2.14

Line of ‘best fit’

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



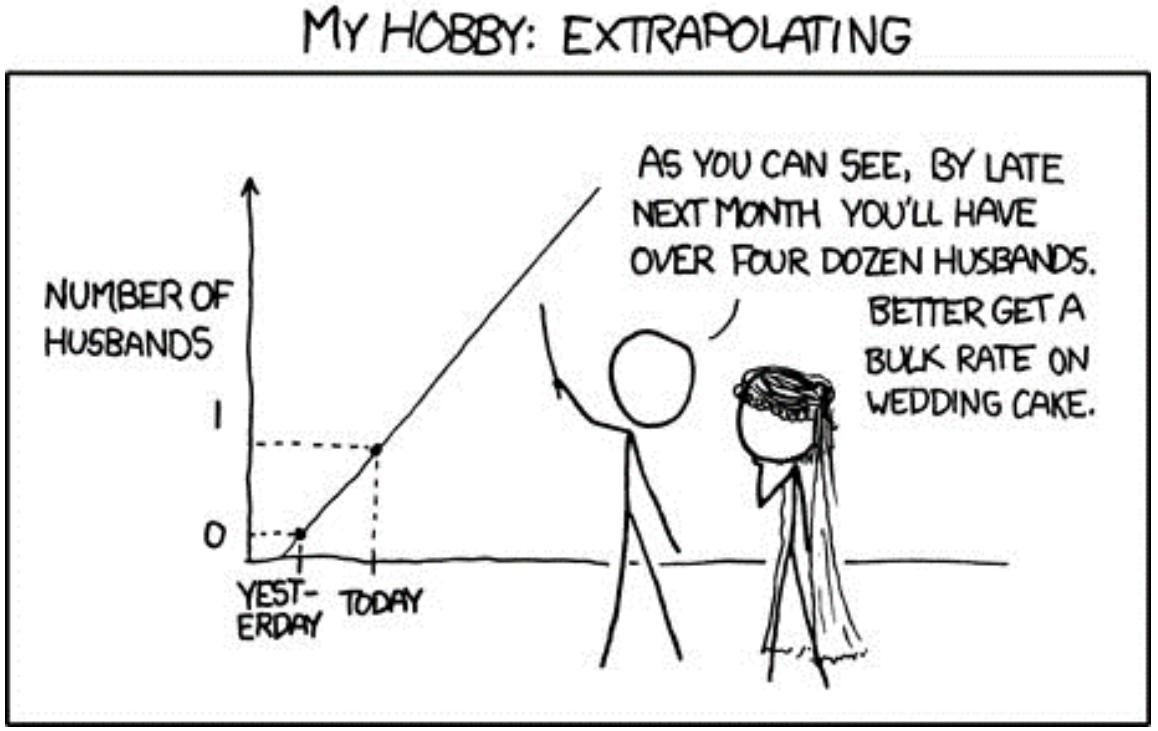
Try to find the line of best fit

Cancer smoking residuals

Cancer obs (y)	Cancer pred ($\hat{y}$)	Residuals $(y - \hat{y})$	Residuals^2 $(y - \hat{y})^2$
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

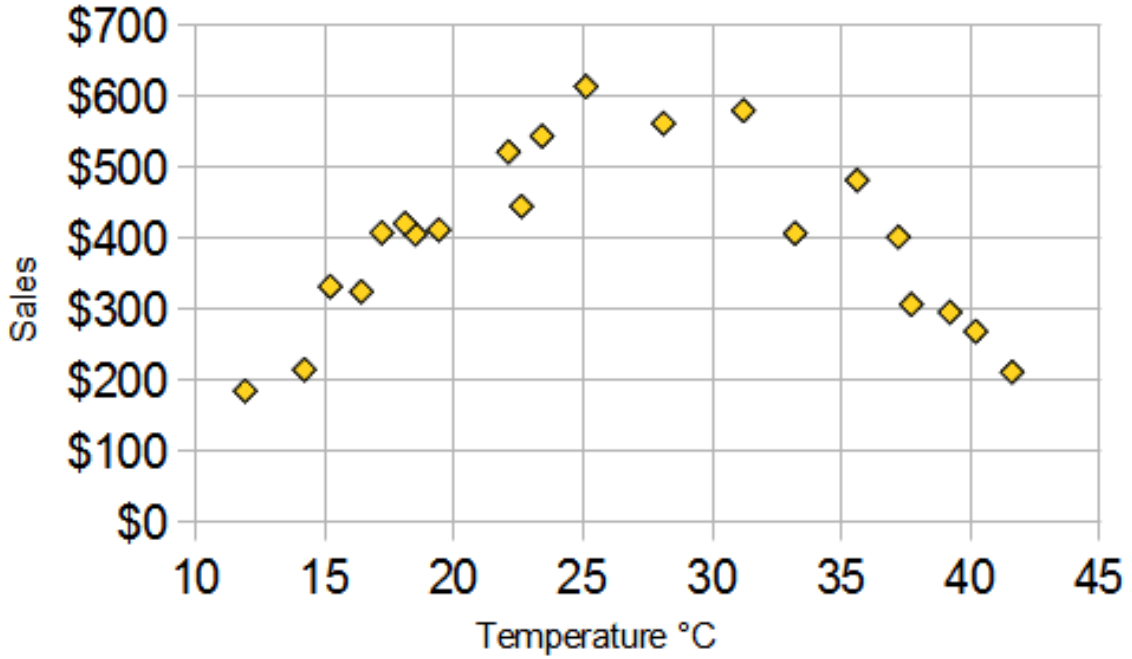
Regression caution # 1

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



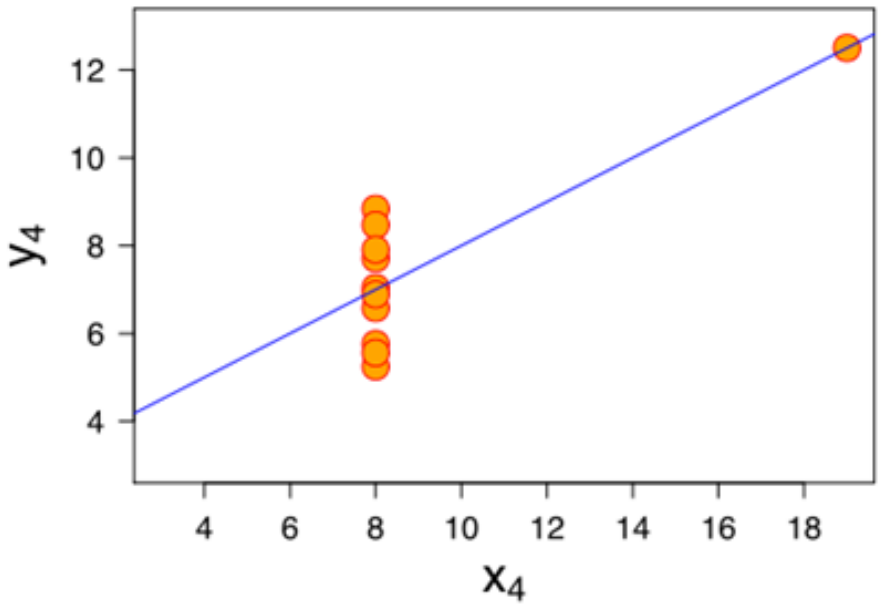
Regression caution # 2

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.



Calculating regression lines in R

```
# download the smoking data
> download_class_data("smoking_cancer.Rda")

# create a scatter plot and calculate the correlation
> plot(smoking$CIG, smoking$LUNG)

# fit a regression model
> lm_fit <- lm(smoking$LUNG ~ smoking$CIG)

# examine the a and b coefficients
> coef(lm_fit)

# add the 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

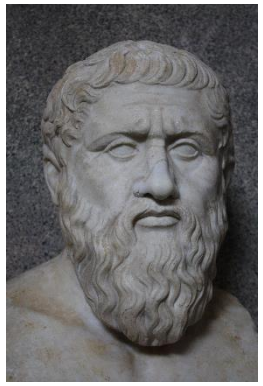
In **linear regression** we fit a line to the data, called the **regression line**

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

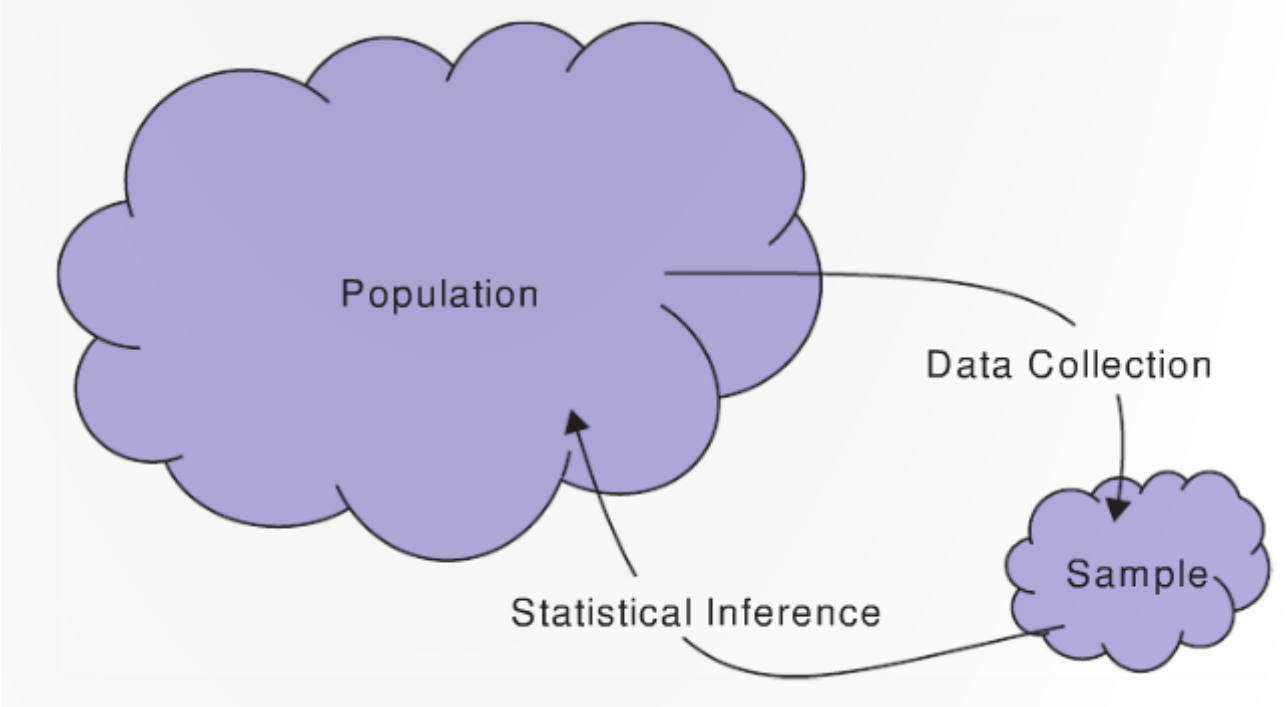
The **residual** is the difference between an observed (y_i) and a predicted value ($\hat{y}_i$) of the response variable

- The regression line minimizes the sum of squared residuals

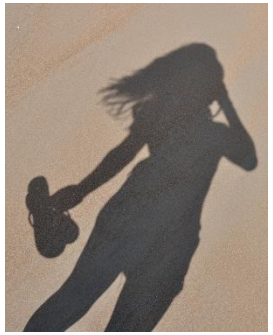
Any last questions about descriptive statistics?



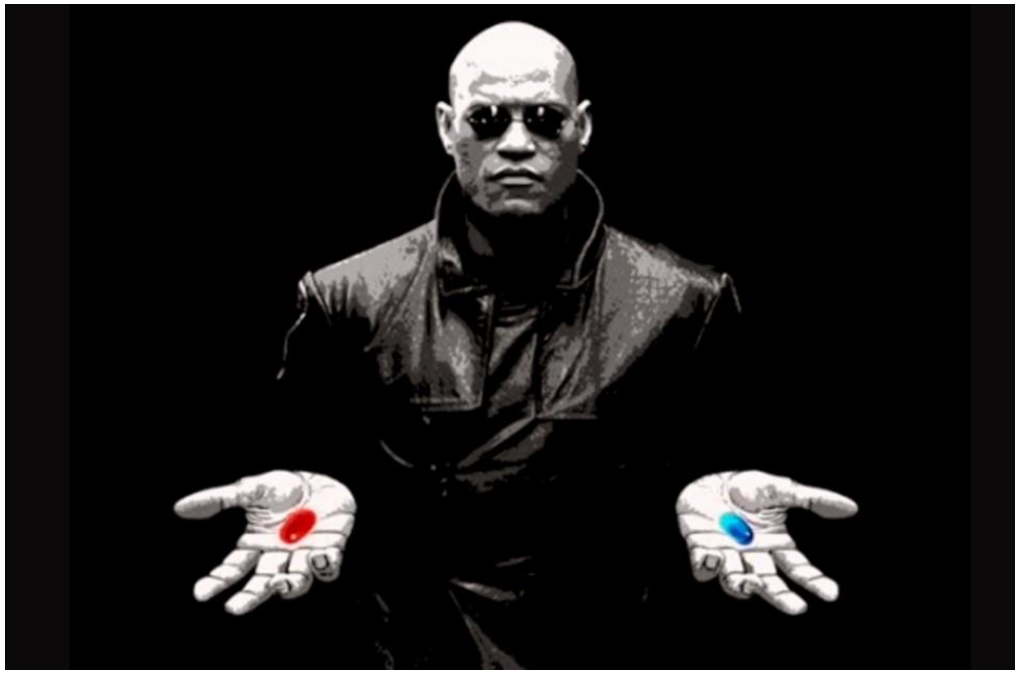
$\pi, \mu, \sigma, \rho, \beta$



$\hat{p}, \bar{x}, s, r, b$



Any last questions about descriptive statistics?



Bias and sampling distributions

Where do samples/data come from?

Example: sampling 100 sprinkles

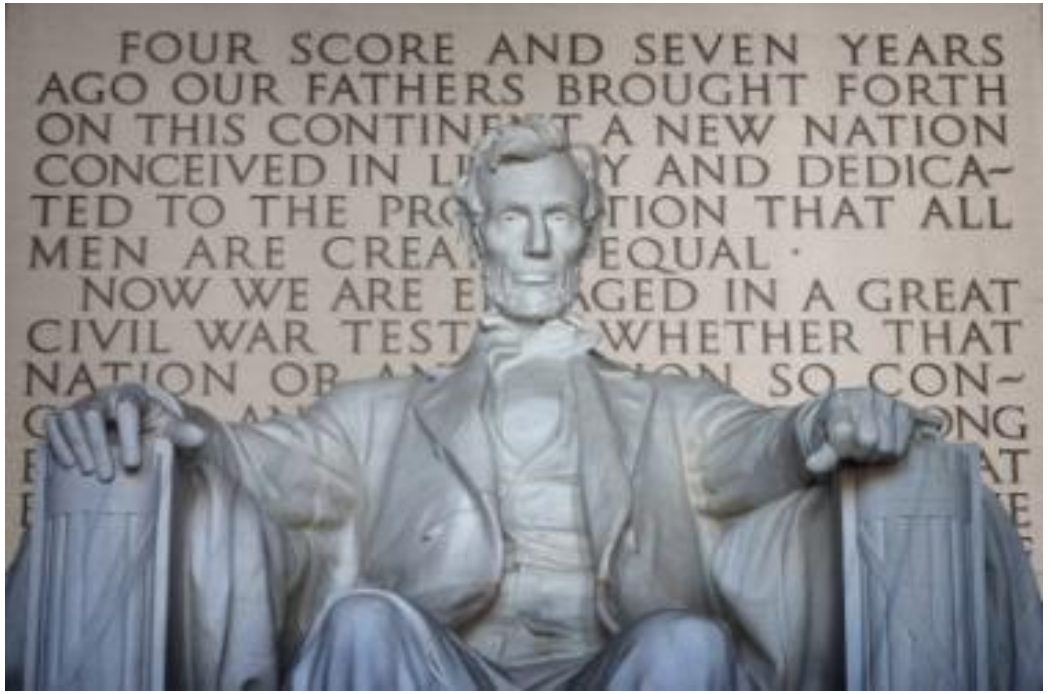
1	orange
2	red
3	green
4	white
5	white
6	white
7	white
8	white
9	red

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

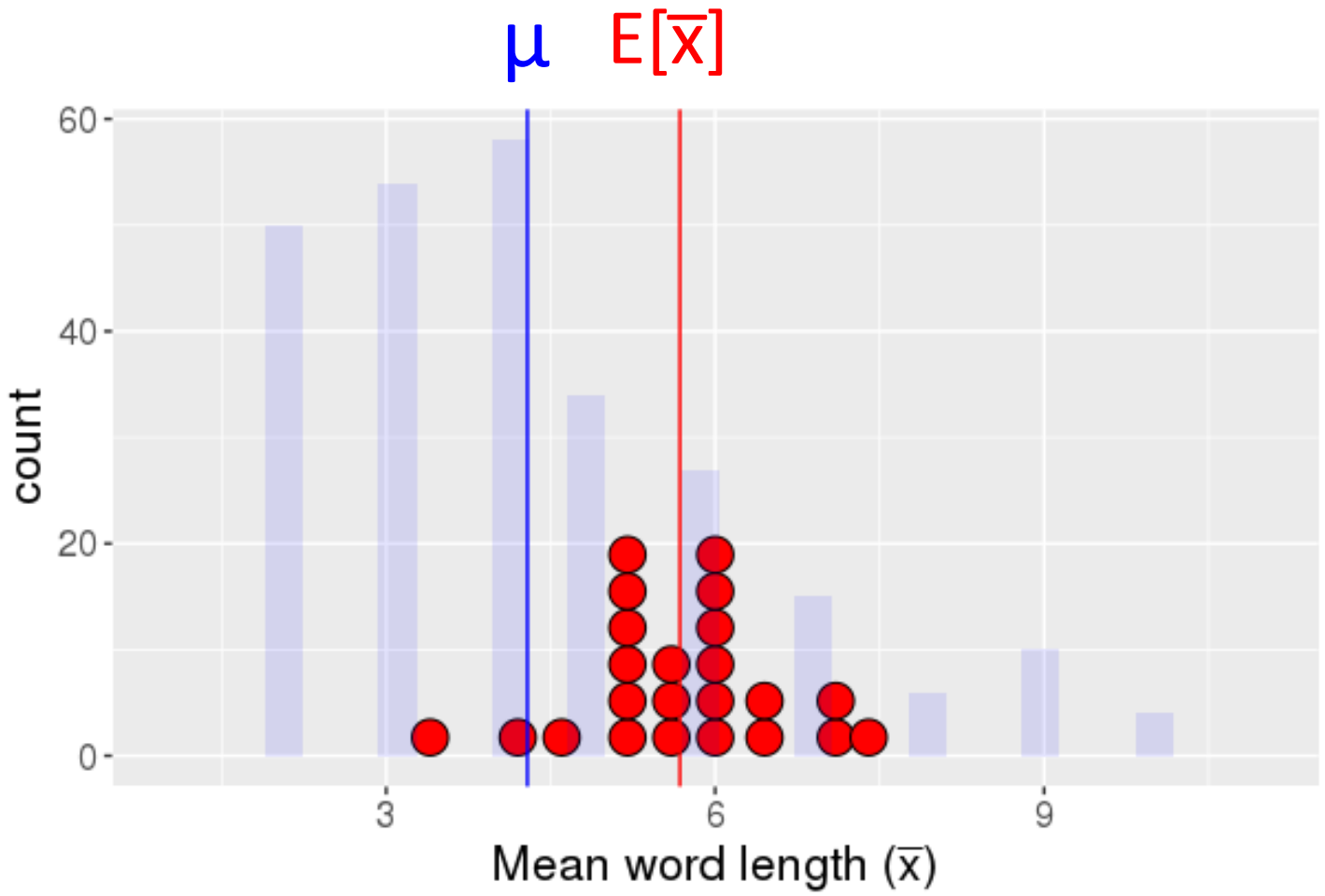
Let’s try some sampling ourselves…

Fill out the worksheet where you need to randomly sample 10 words from the Gettysburg address

Report the mean of the 10 words at:

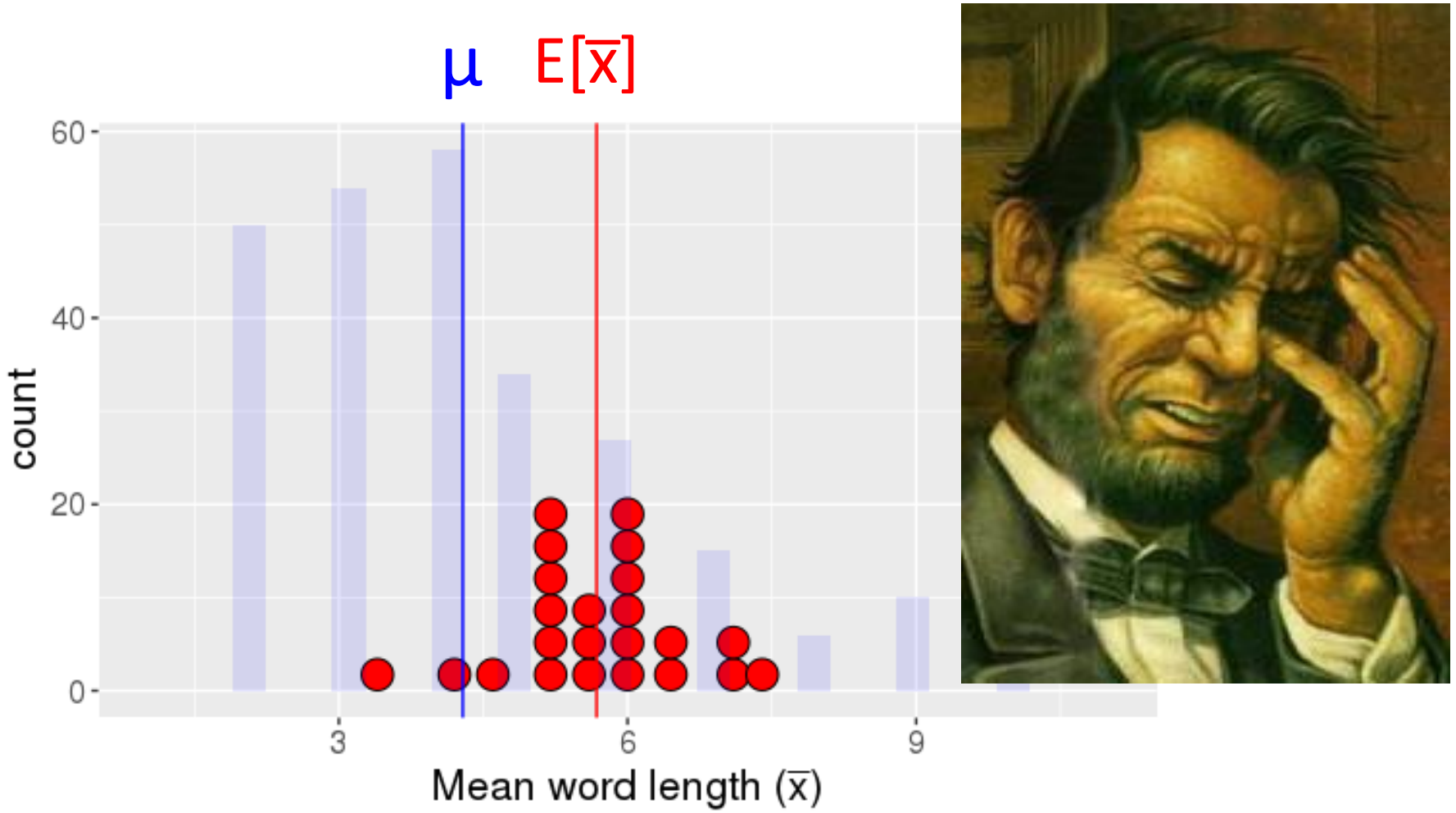


Gettysburg address, mean word length



Gettysburg address, mean word length

Observations?

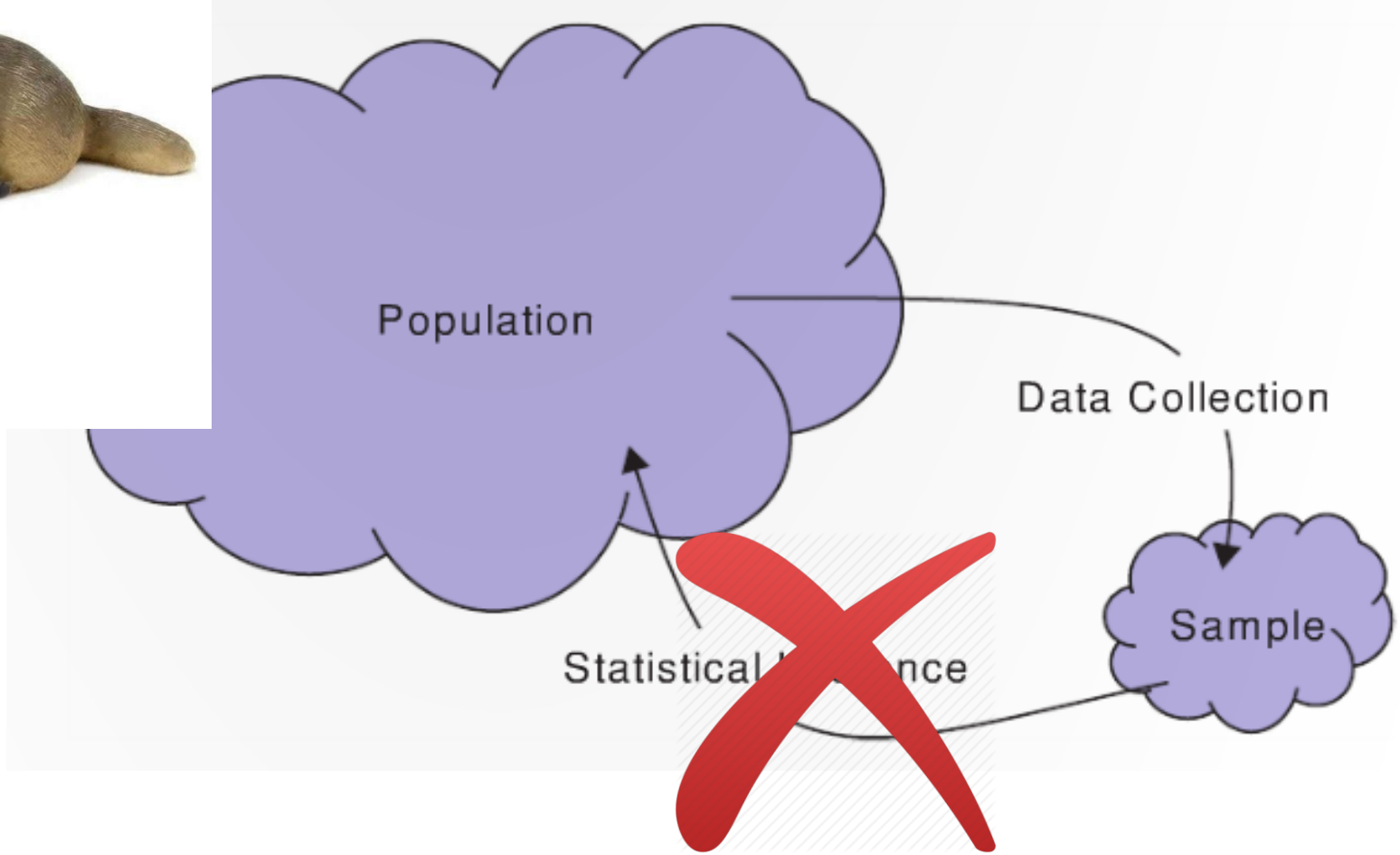


Other types of bias

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

Statistical bias

μ



$\bar{x}$

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?

A poll for the Truman/Dewey election that randomly chose 6,000 people from all citizens 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 administered by the campus's offices. A total of 1,920 surveys were completed, yielding a response rate of 24%.

Alumni Survey Results

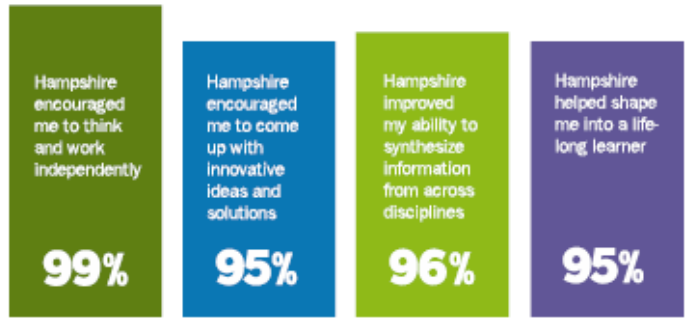
Hampshire College

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



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.

"

Bias or No Bias?

Yelp reviews of restaurants?

An anonymous survey randomly select 6,000 people and ask them 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 they asked do you support “stricter gun control laws”?

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

Second poll they do you support “stricter gun laws”?

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

How could this affect the newspaper headlines?

- “Majority of Americans ***oppose*** stricter gun control laws” vs.
- “Majority of Americans ***support*** stricter gun laws”

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 people left-handed in the US, randomly sampling Yale students might be good enough.

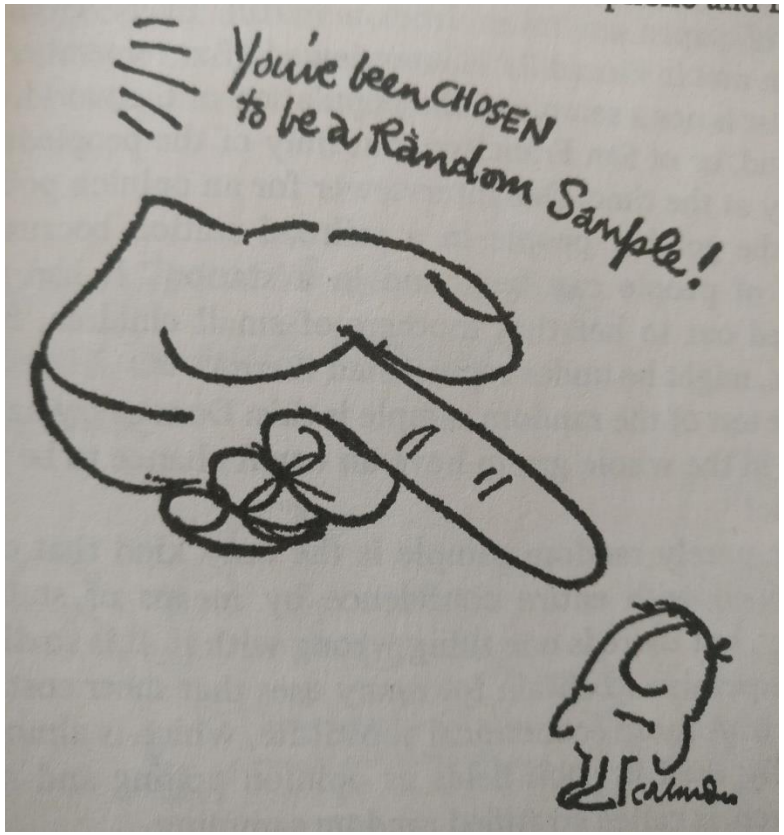
Need to think carefully to avoid bias!

As mentioned last class, statistics requires thought!

Use your own reasoning:

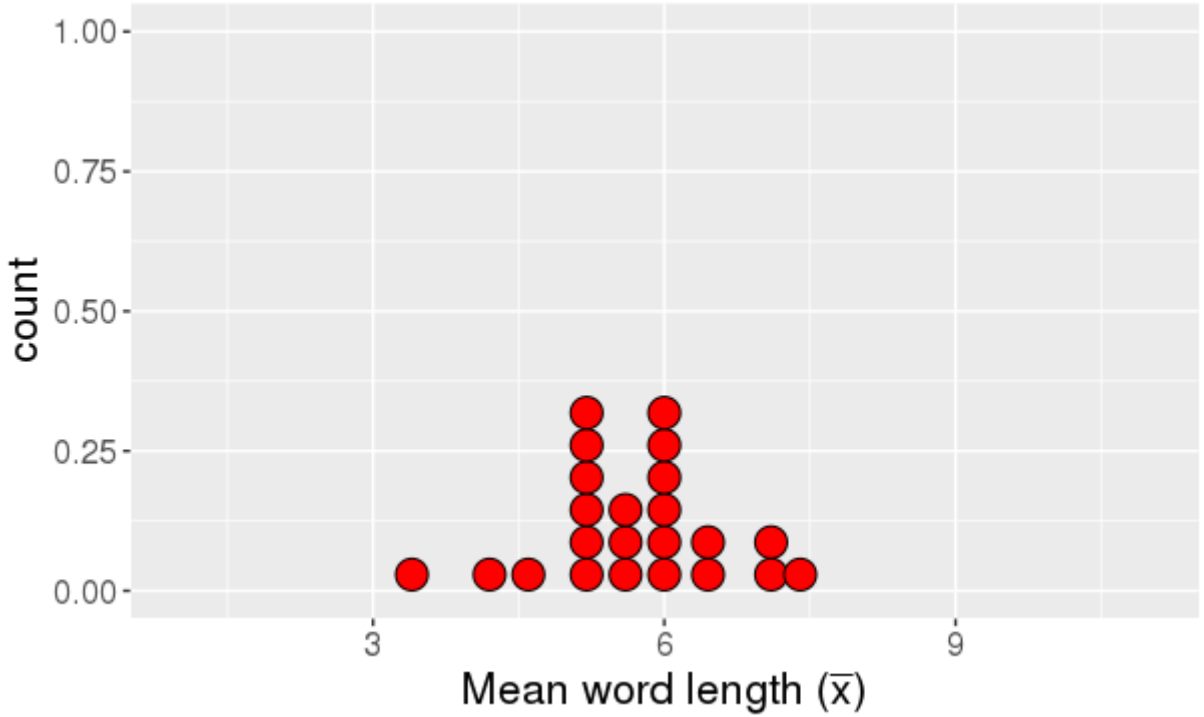
- What is the population I am interested in?
- Does the sample reflect the population of interest?
- Be your own worst critic!

Questions about statistical bias?



For our distribution of Gettysburg word lengths…

Q: What does each case that is plotted correspond to?



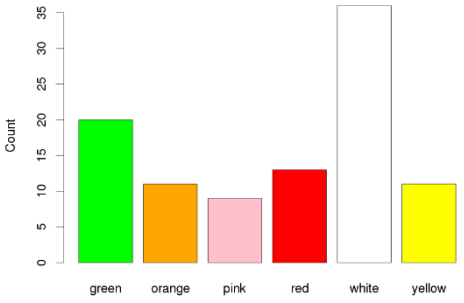
A: The mean length of 10 words ($\bar{x}$)

i.e., each point in our **distribution** is a statistic!

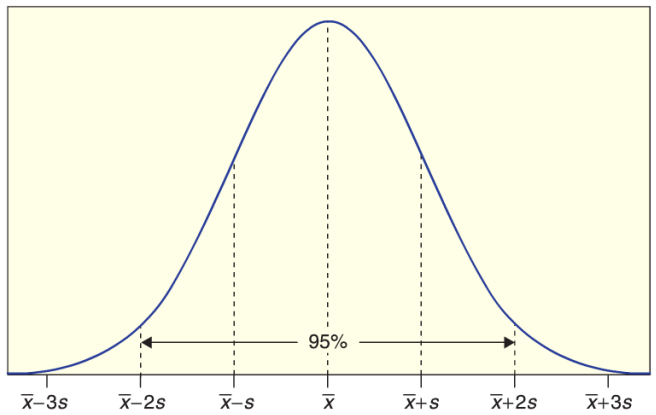
Sampling distribution

A **sampling distribution** is the distribution of sample statistics computed for different samples of the same size (n) from the same population

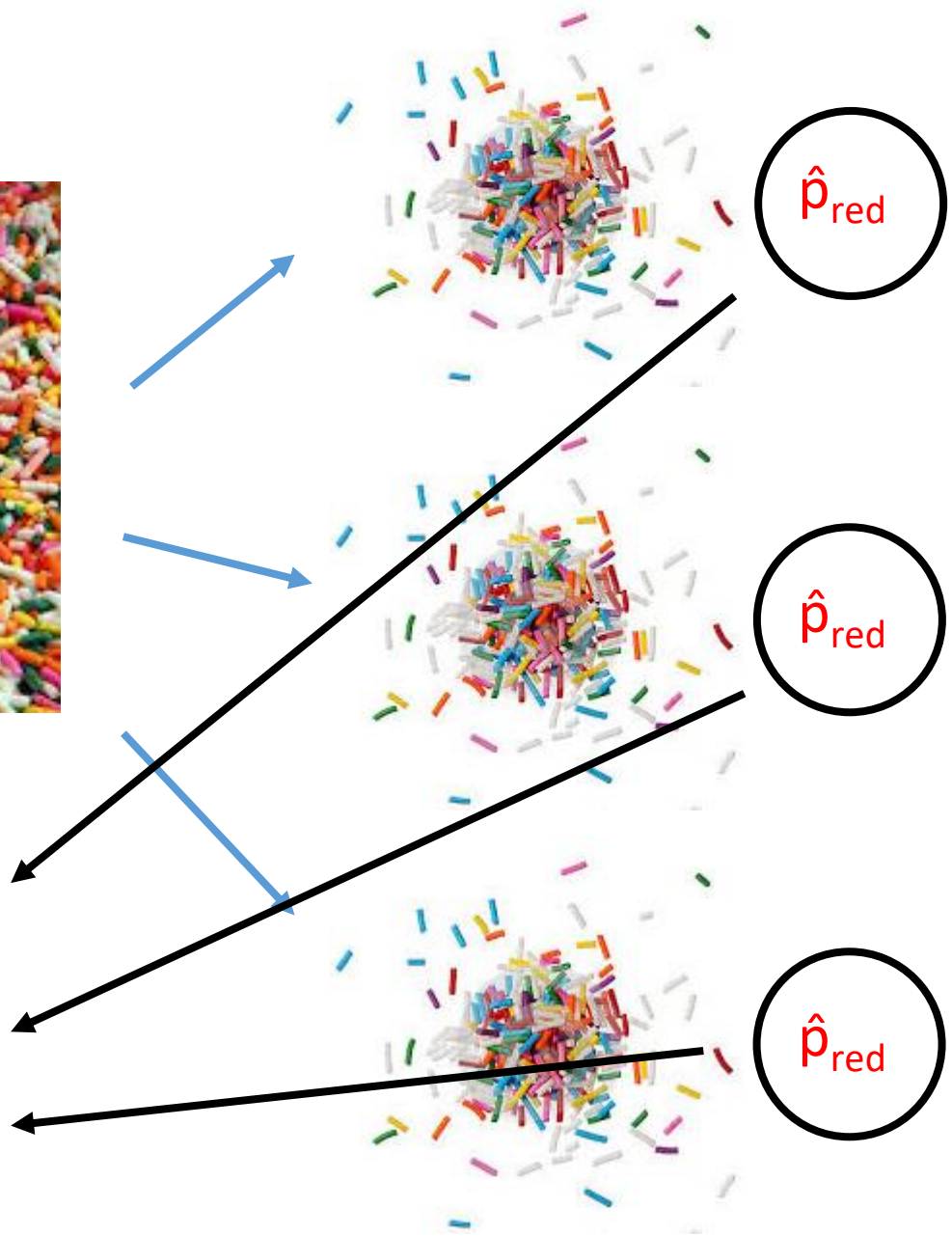
A sampling distribution shows us how the sample statistic varies from sample to sample



π_{red}



Sampling distribution!



Next class

Sampling distributions (in R) and confidence intervals…

Homework 3 has been posted

- Use the link on Canvas to access homework 3 on R Studio Cloud
- Due on Gradescope at 11:30pm on Sunday February 9th