

Sampling, bias and sampling distributions

Overview

Review of simple linear regression and R regression practice

Sampling and bias

Sampling distributions

Announcements

Homework 3 has been posted

- It is due on Gradescope on Sunday February 11th at 11pm
- [SDS100::download_homework\(3\)](#)
- Be sure to mark all pages for each question
 - Points will be taken off if pages are not marked correctly!

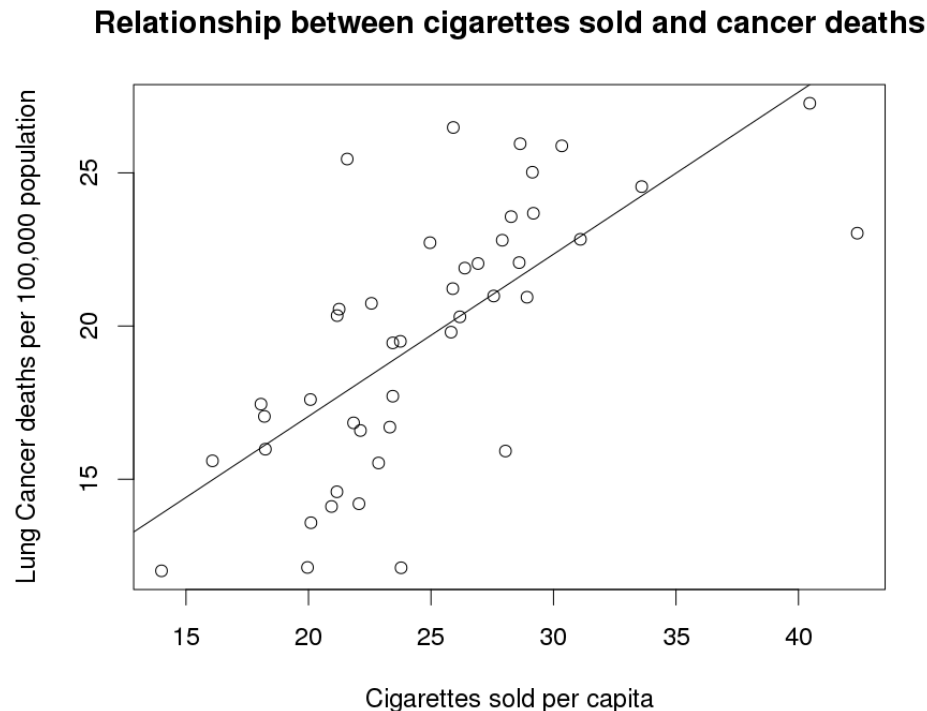
Questions about anything?

Review of linear regression

Review: Linear regression

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

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



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

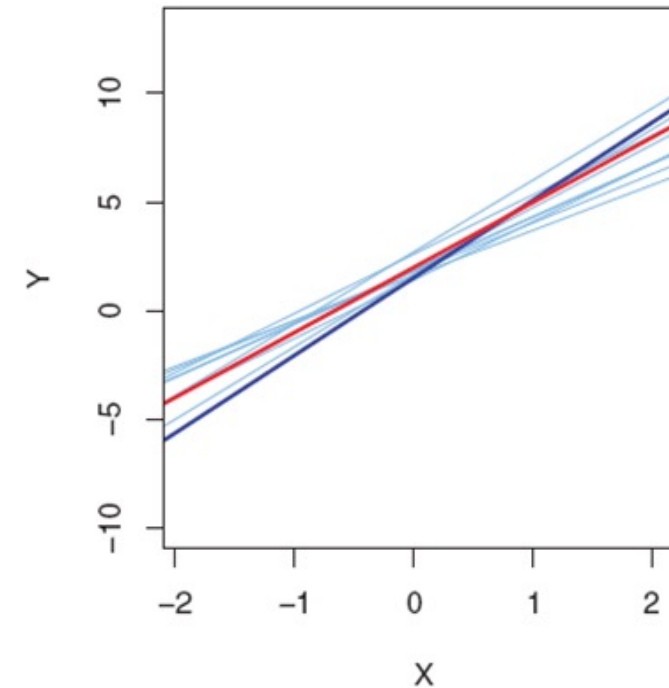
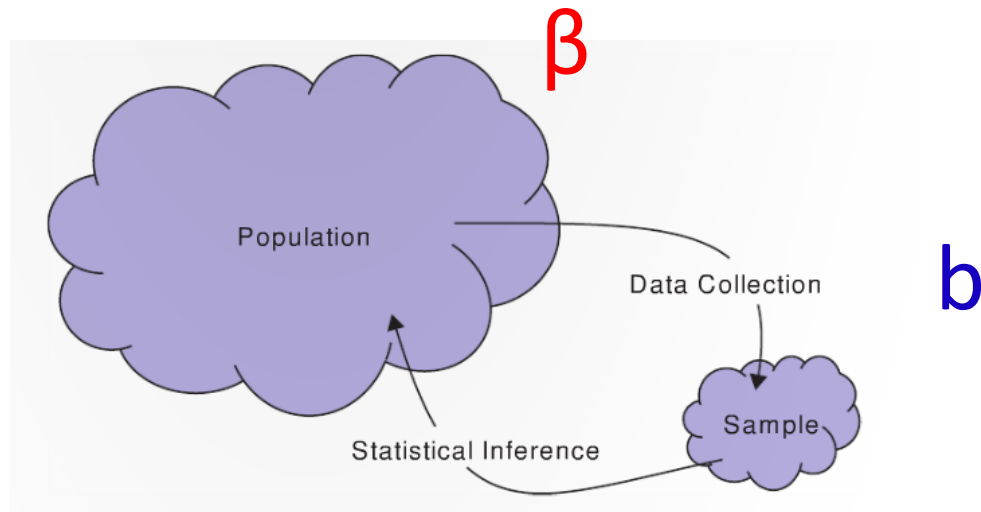
$$a = 6.47 \quad b = 0.53$$

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

Review: 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**



Review: Residuals and the least squares line

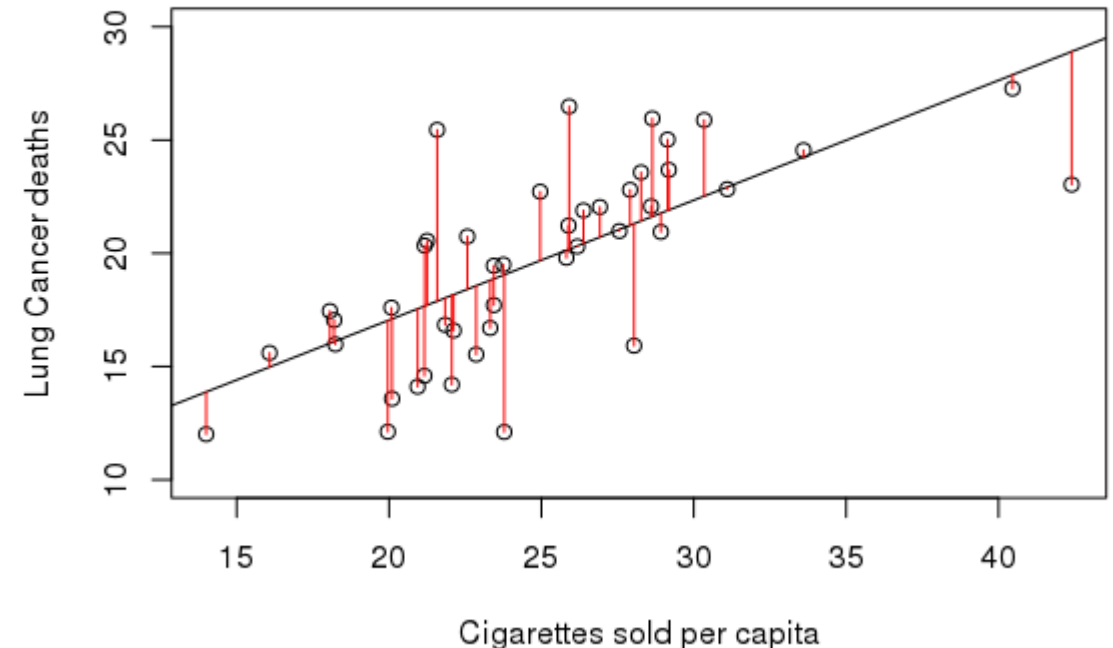
The **residual** is the difference between an observed (y_i) and a predicted value (\hat{y}_i) of the response variable

- $e_i = y_i - \hat{y}_i$

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

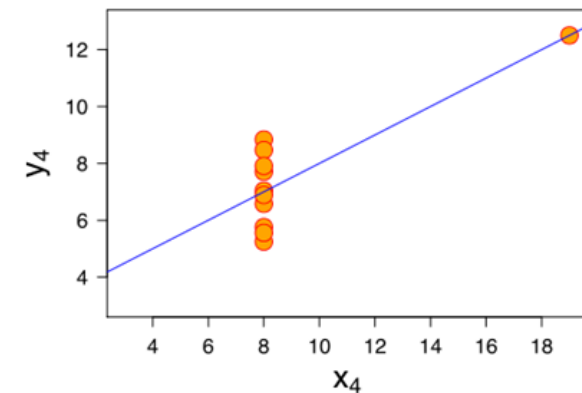
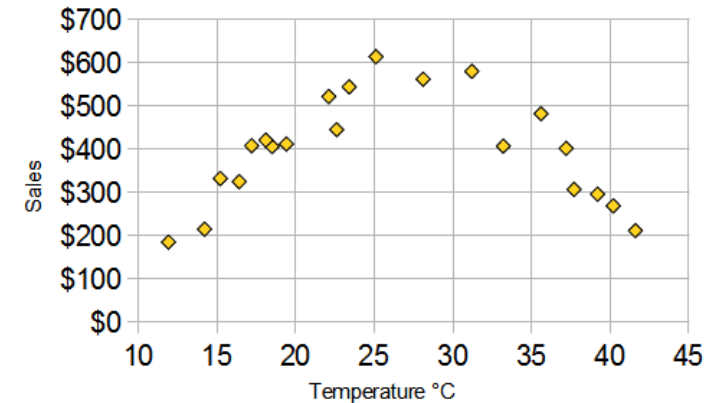
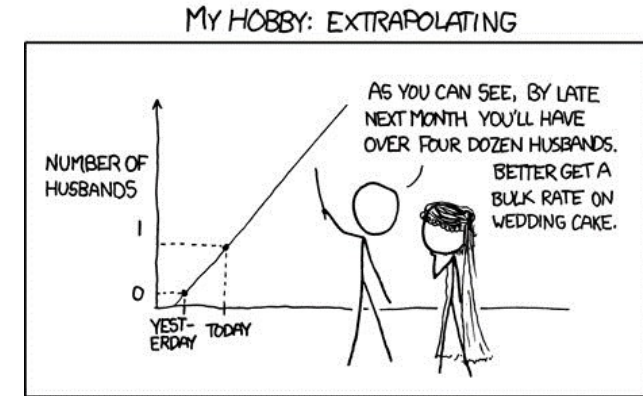
[Find the line of best fit](#)

Relationship between cigarettes sold and cancer deaths

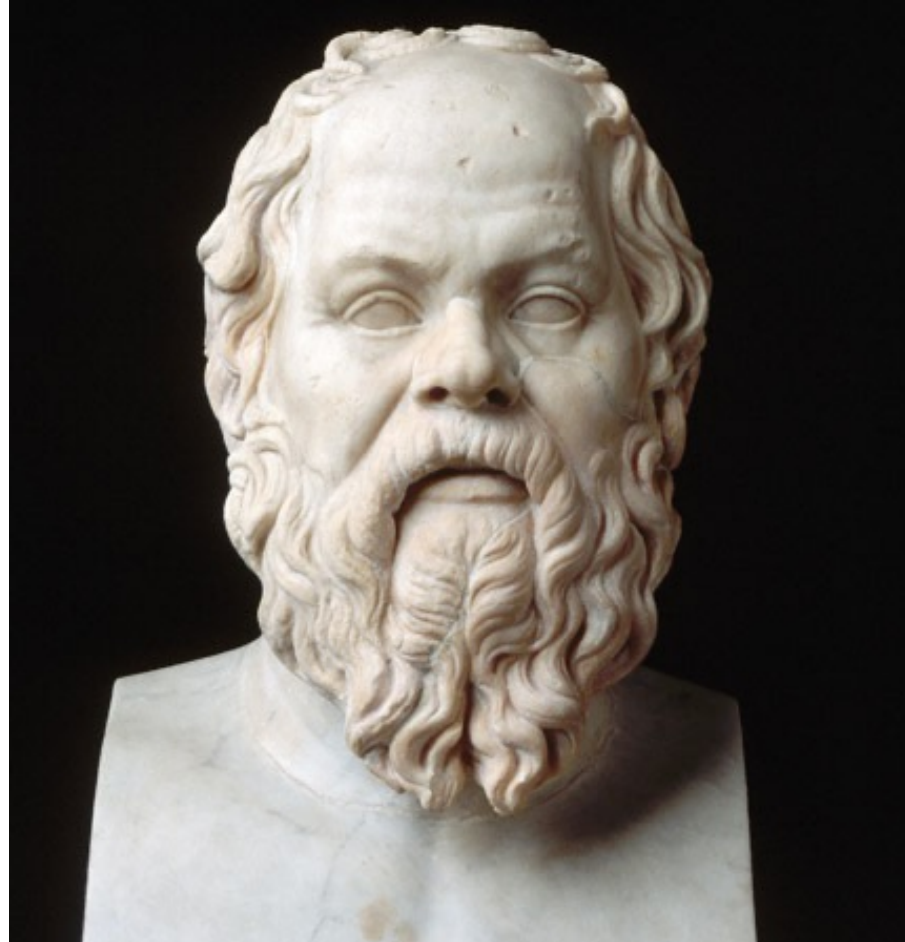


Regression cautions

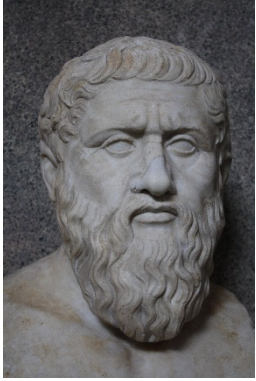
1. Avoid trying to apply the regression line to predict values far from those that were used to create the line.
2. Plot the data! Regression lines are only appropriate when there is a linear trend in the data.
3. Be aware of outliers – they can have an huge effect on the regression line.



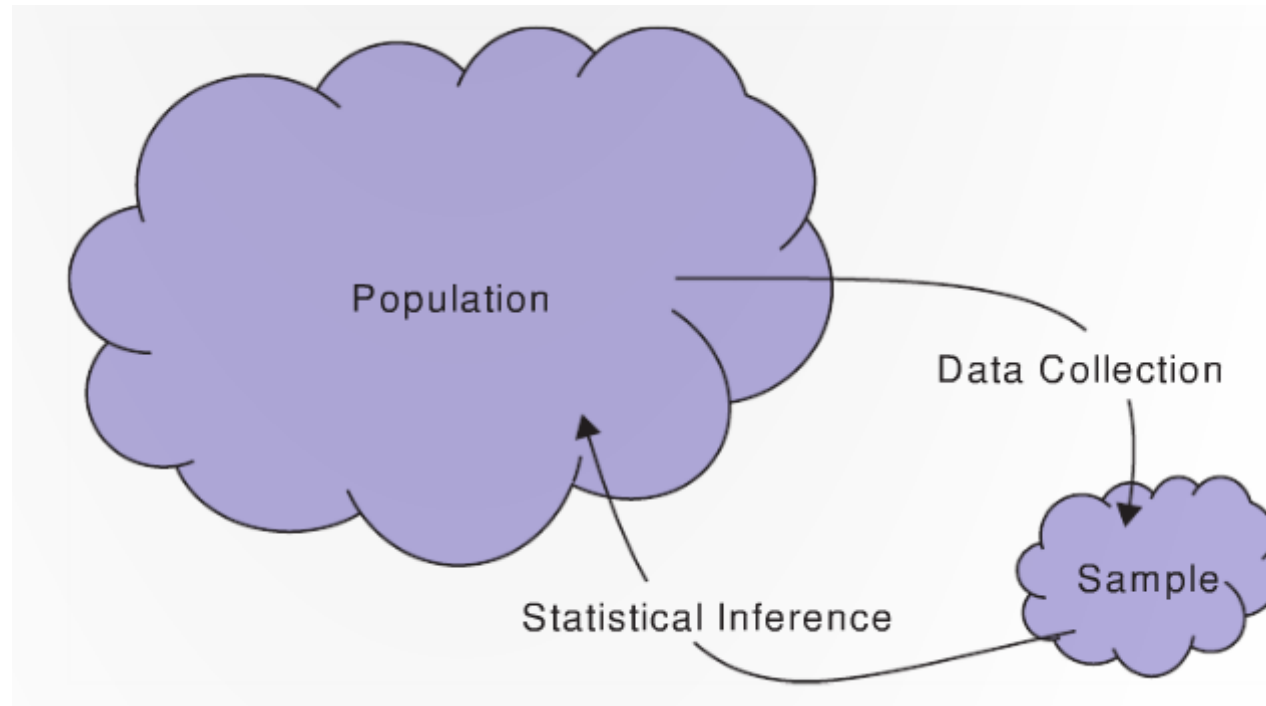
Socratic method to review descriptive statistics



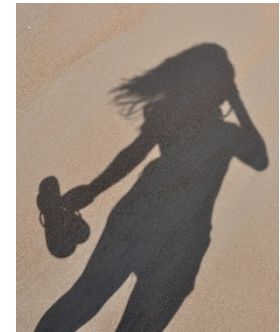
Parameters and statistics



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



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



THE TRUTH IS OUT THERE



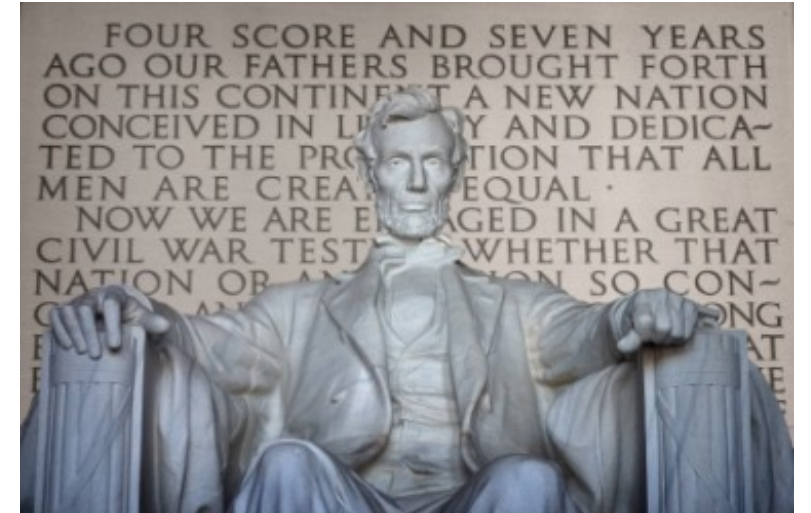
Can you handle the Truth[®]?



Questions?

Sampling

Where do samples/data come from?



Q: What symbol do we use to denote the sample size?

- n

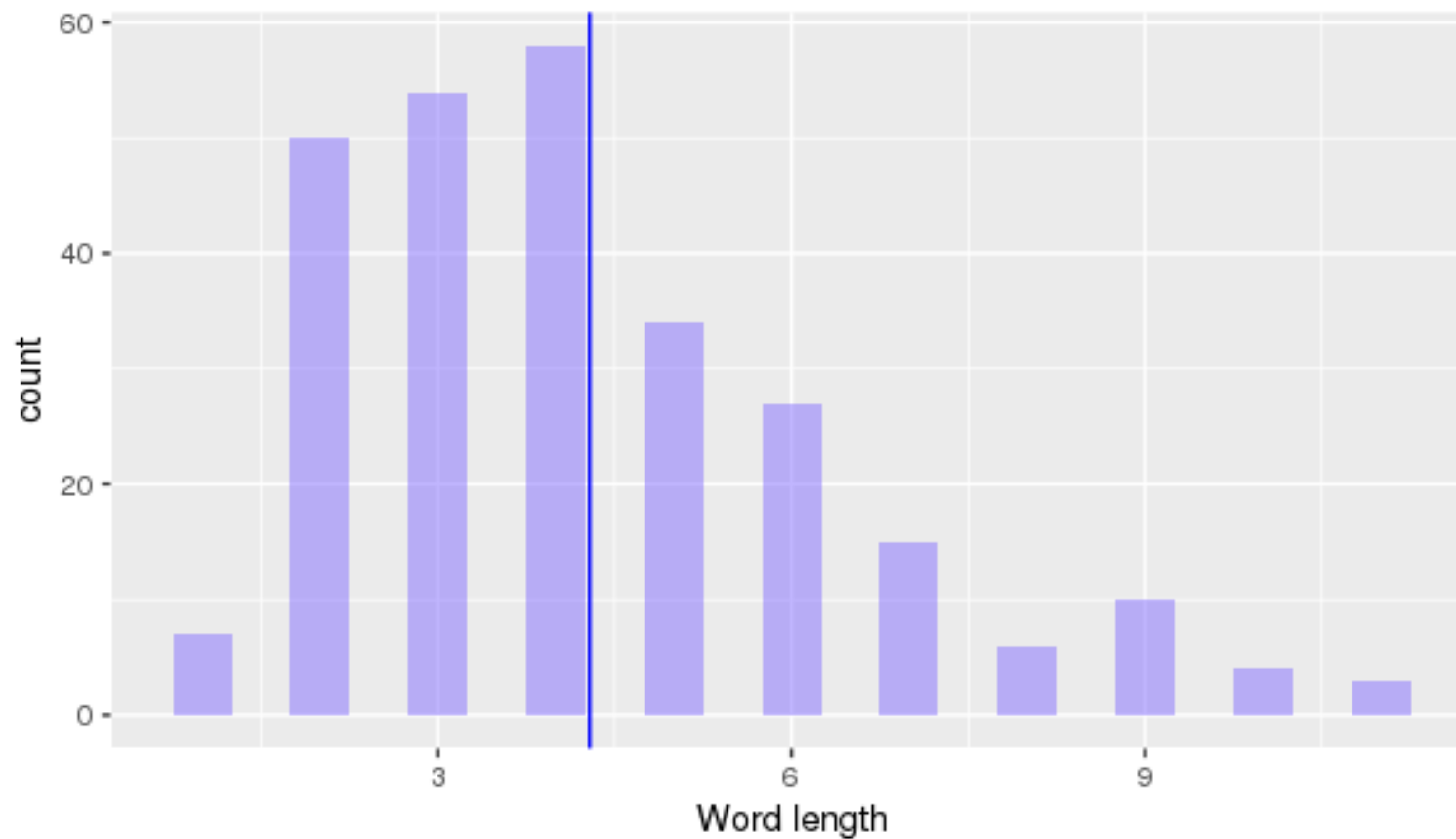
The Gettysburg address has 268 words

Students sampled $n = 10$ words randomly from a print out of the address

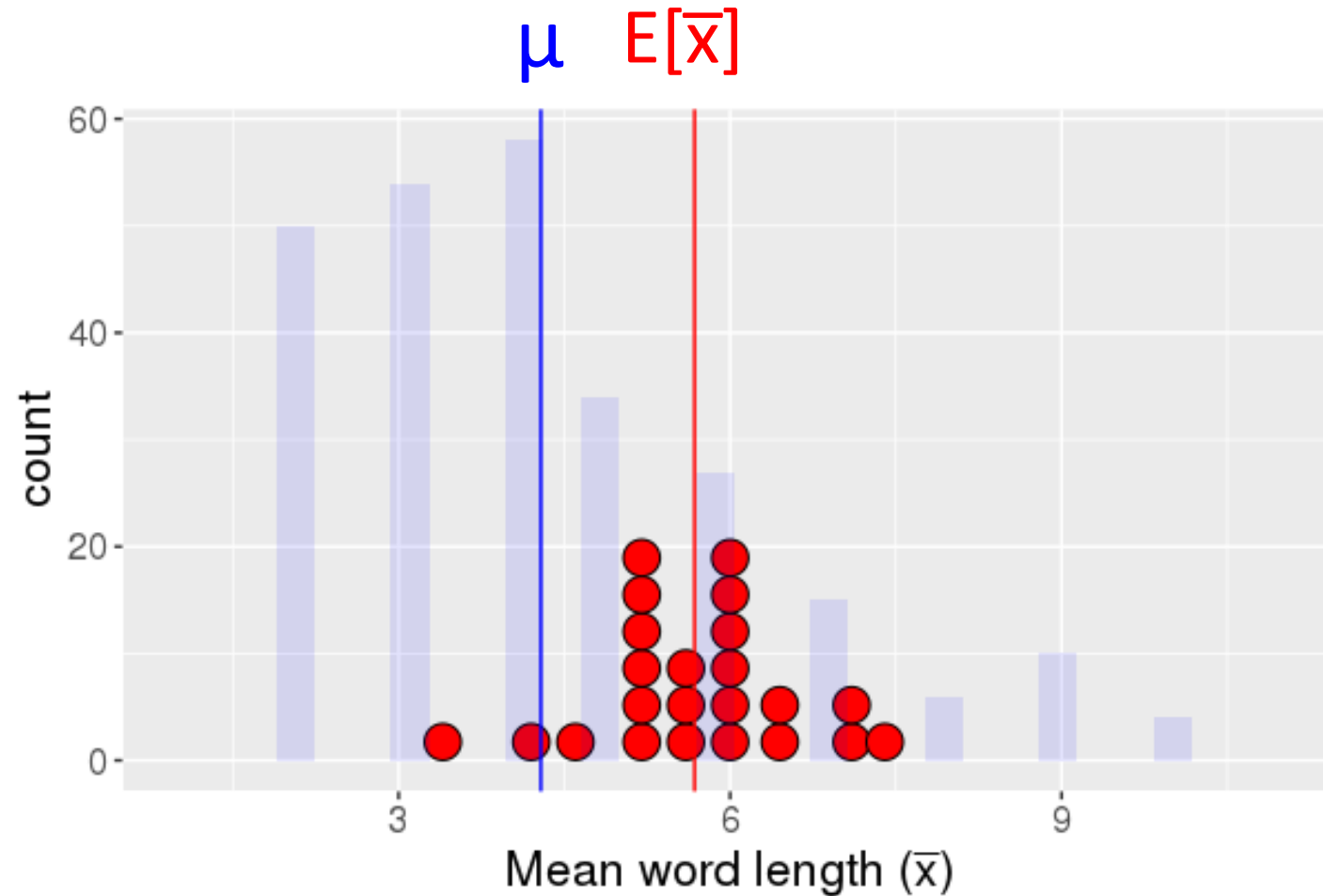
- Did you try this?

Gettysburg address: lengths of 268 words in the population

$$\mu = 4.287$$

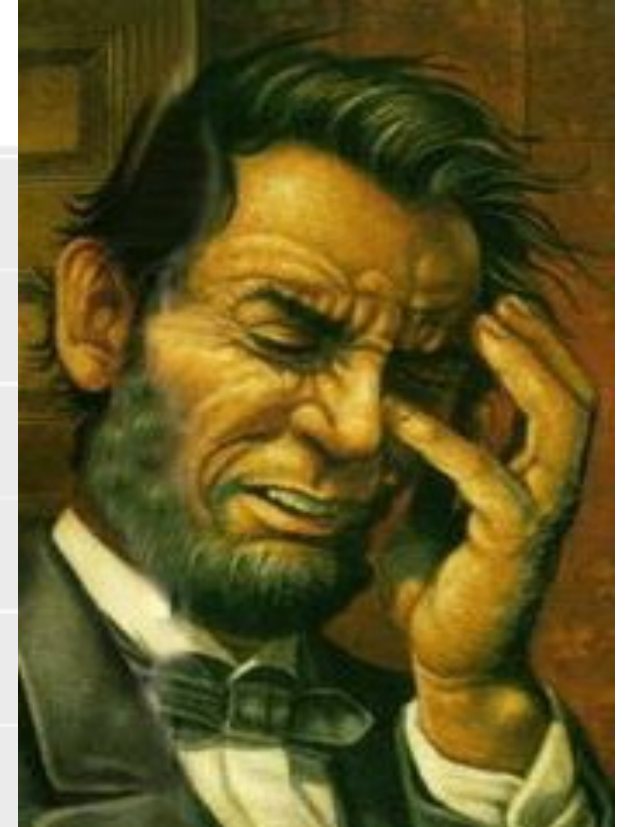
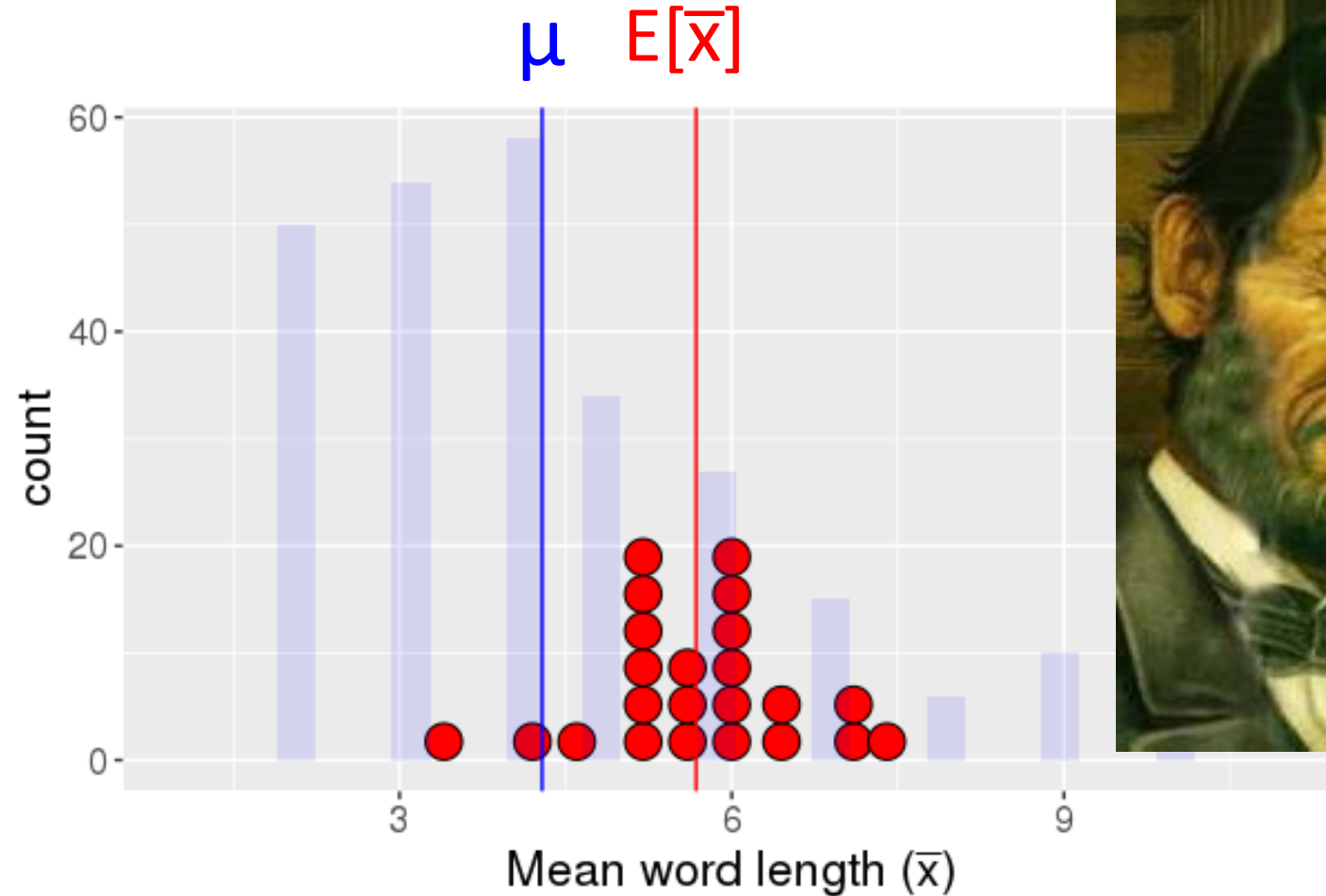


Gettysburg address, mean word length



Gettysburg address, mean word length

Observations?



Let's explore your average word lengths in R...

```
library(SDS100)
```

```
download_class_code(7)
```

Sampling and bias

Bias

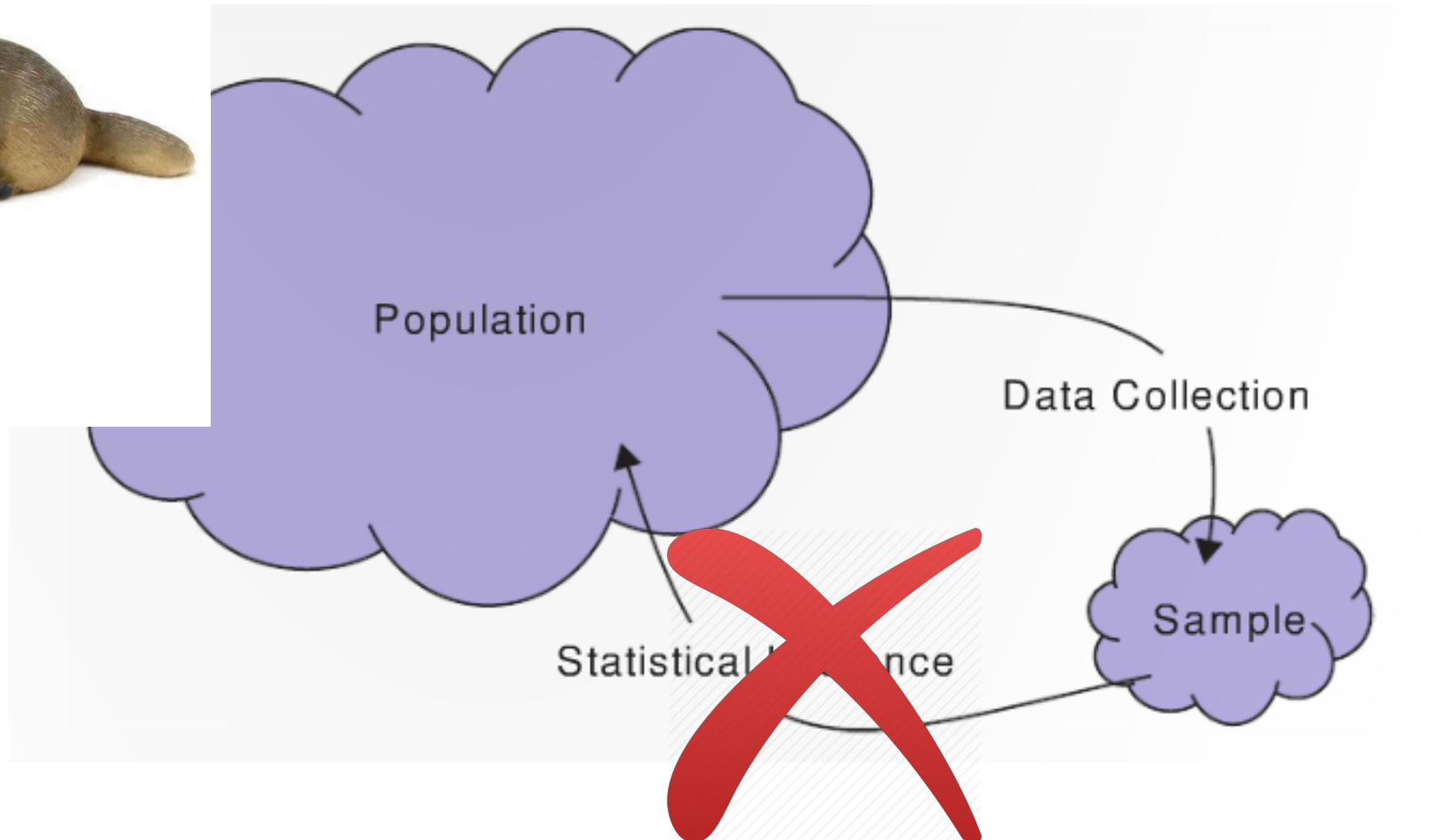
Sampling bias exists when the method of collecting the data causes the sample to inaccurately reflect the population.

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

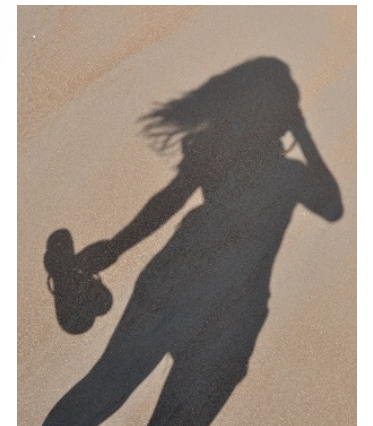
- E.g., $E[\bar{x}] \neq \mu$

Statistical bias

μ



\bar{x}



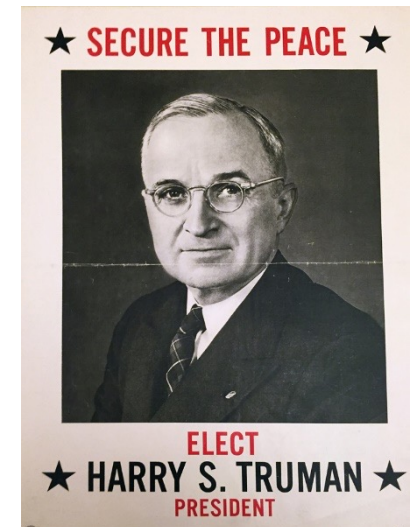
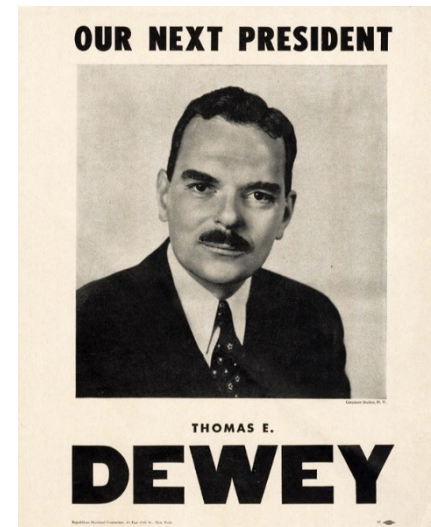
Basic questions for sampling

What is the population?

What is the sample?

Do they differ in a meaningful way?

1948 US election



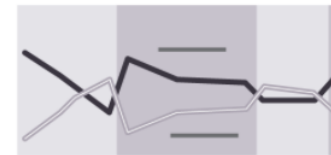
More recent results...

PEW RESEARCH CENTER | MARCH 2, 2021



What 2020's Election Poll Errors Tell Us About the Accuracy of Issue Polling

BY SCOTT KEETER, NICK HATLEY, ARNOLD LAU AND COURTNEY KENNEDY



CNN politics

LIVE TV CNN+  

Here's what pollsters think happened with 2020 election surveys



Analysis by Ariel Edwards-Levy, CNN
Updated 3:27 PM ET, Thu May 13, 2021



RESEARCH NEWS

Pre-election polls in 2020 had the largest errors in 40 years

Jul. 19, 2021, 8:00 AM

Share    

 Pre-election polls in 2020 had the largest errors in 40 years

Watch later 

FiveThirtyEight

Politics Sports Science Podcasts Video

FEB. 23, 2021, AT 6:00 AM

Why Was The National Polling Environment So Off In 2020?

A look at how congressional polls, just like those for president, missed in 2020.



 TheUpshot

POLITICAL CALCULUS

A 2016 Review: Why Key State Polls Were Wrong About Trump

      192



**Bias or
no bias?**



Amazon reviews of products?

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

A professor asks his class, by a show of hands, how many people have watched the pre-recorded class videos prior to coming to class?

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



How many people wash their hands after using the restroom...?

- a. A study asked 6,000 randomly selected people if they wash their hands after using the restroom.
- b. A study from Harris Interactive collected data by standing in public restrooms and pretending to comb their hair or put on make-up and observed whether 6,000 patrons washed their hands.



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 asked: do you support “stricter gun laws”?

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



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



Hot sauce analogy



How do we select a random sample?

1. List everyone in the population
2. Select randomly by:
 - a. Mechanical means:
 - Flip coins
 - Pull balls from well mixed bins
 - Deal out shuffled cards, etc.
 - b. Use computer programs



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!

Statistics requires thought!

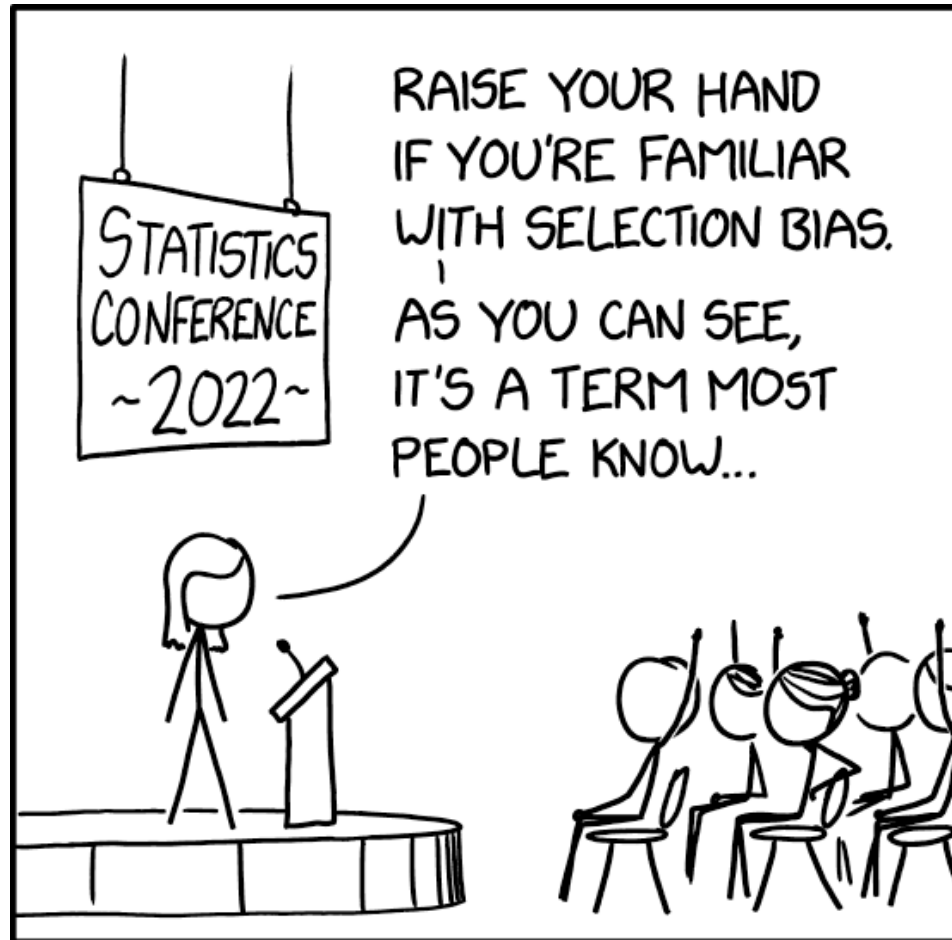
Use your own reasoning:

- What is the population I am interested in?

- Does the sample reflect the population of interest?

- Be your own worst critic!

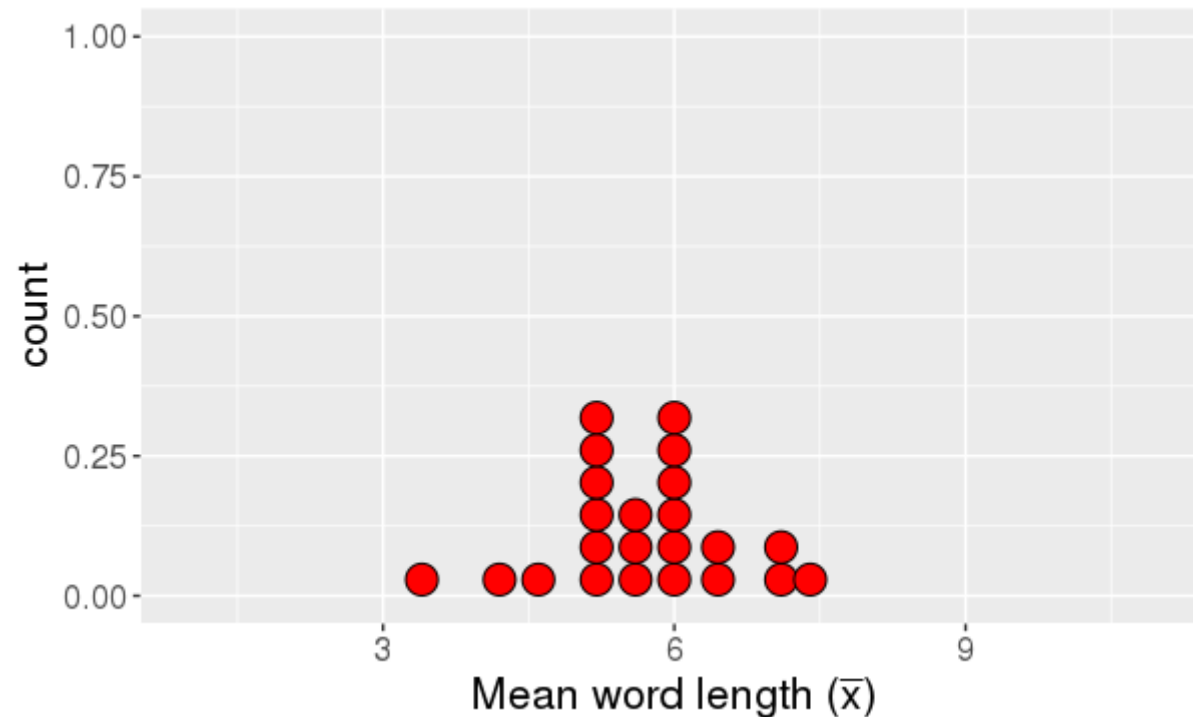
Questions about statistical bias?



Sampling distributions

Recall for our distribution of Gettysburg word lengths...

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

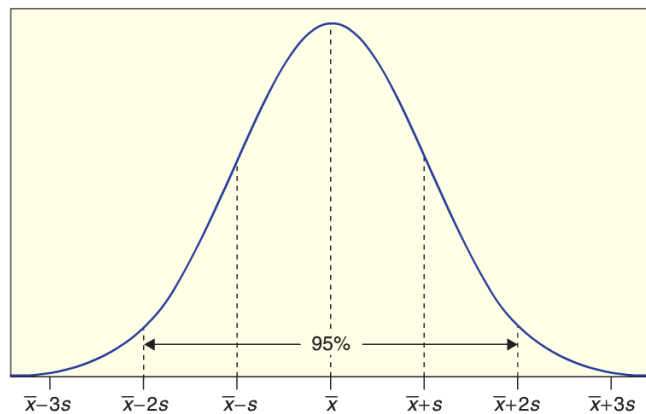
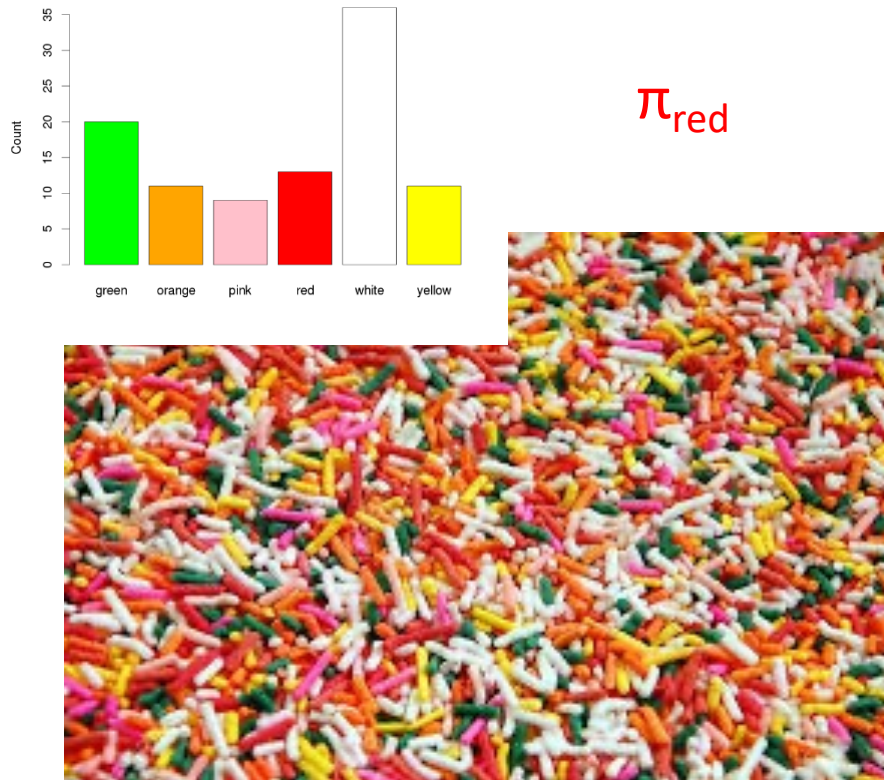


A: The mean length of 10 words (\bar{x})
i.e., each point in our **distribution** is a statistic!

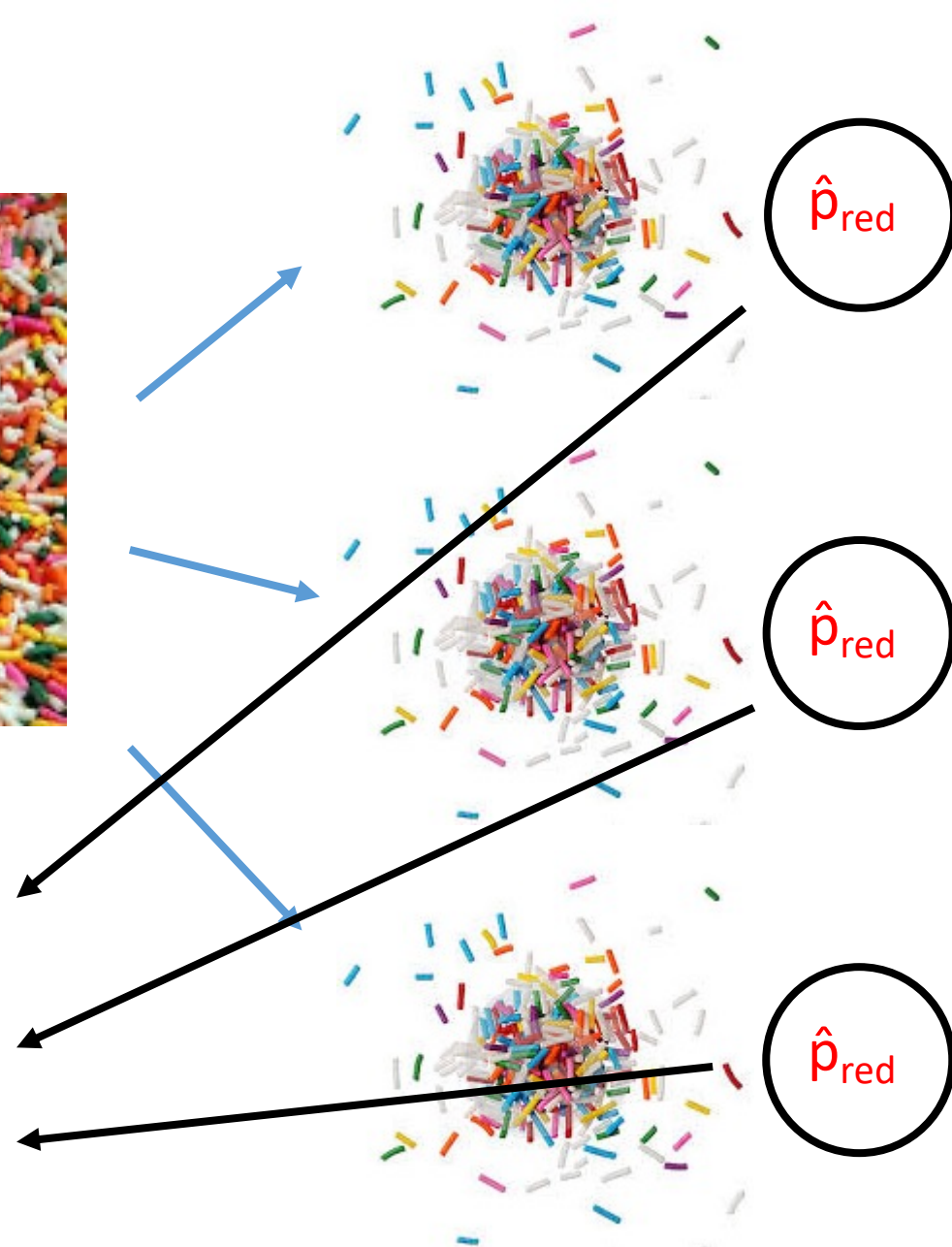
Sampling distribution

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

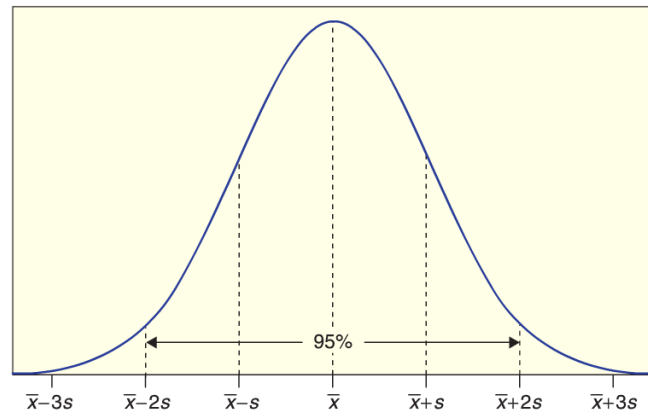
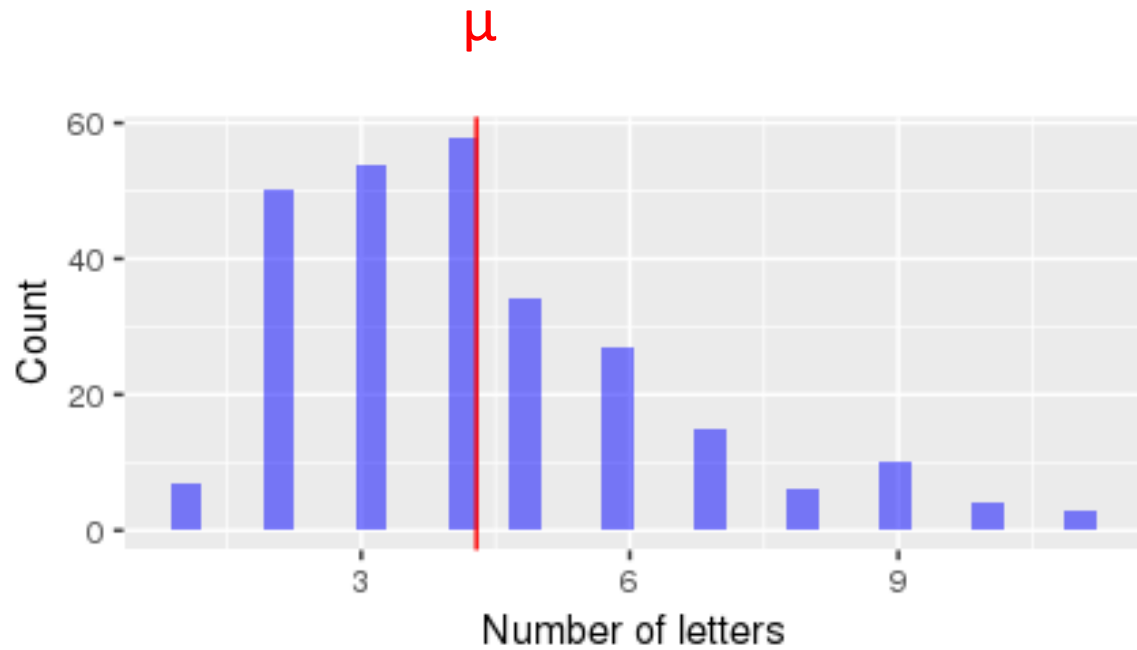
A sampling distribution shows us how the sample statistics vary from sample to sample



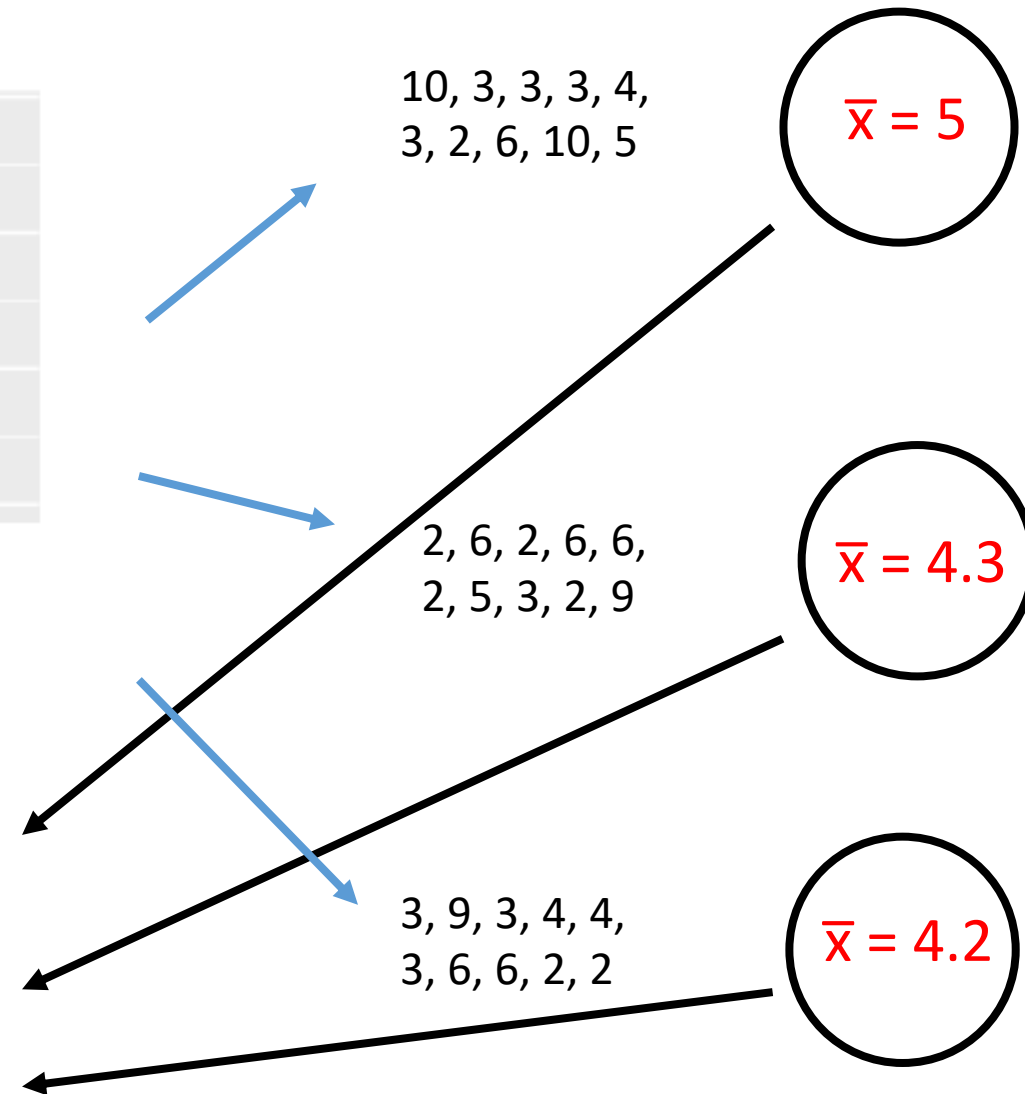
Sampling distribution!



Gettysburg address word length sampling distribution



Sampling distribution!



[Gettysburg sampling distribution app](#)

Let's try it in R!