Bias and sampling distributions

Overview

Questions about worksheet 4

Very quick review of correlation cautions

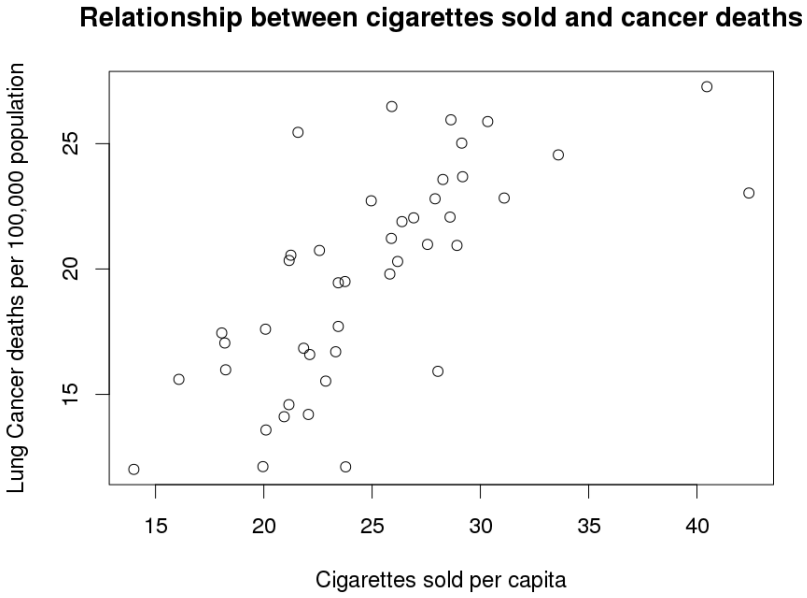
Sampling and bias

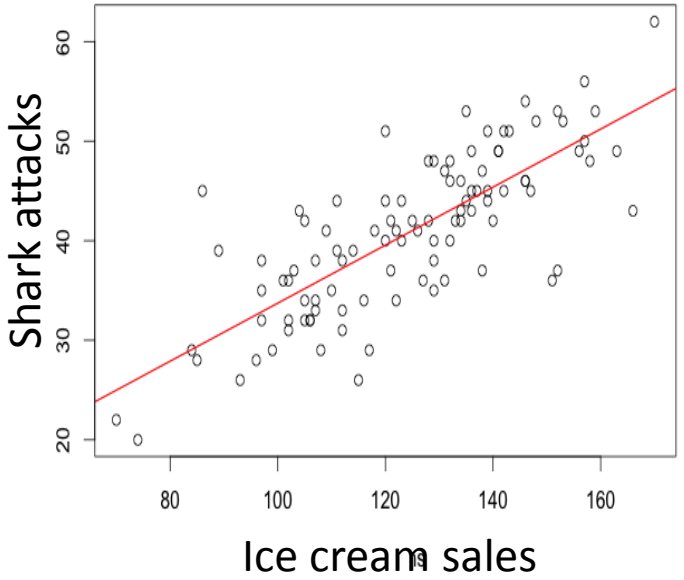
Sampling distributions

Any questions about worksheet 4?

Review of correlation cautions

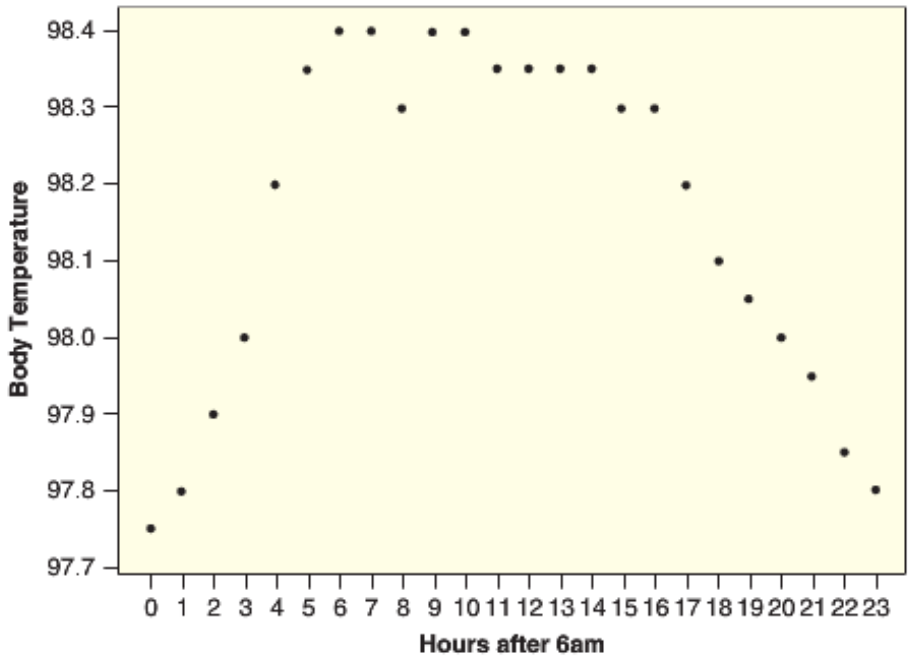
1. Does correlation ($\rho \neq 0$) imply causation?





Review of correlation cautions

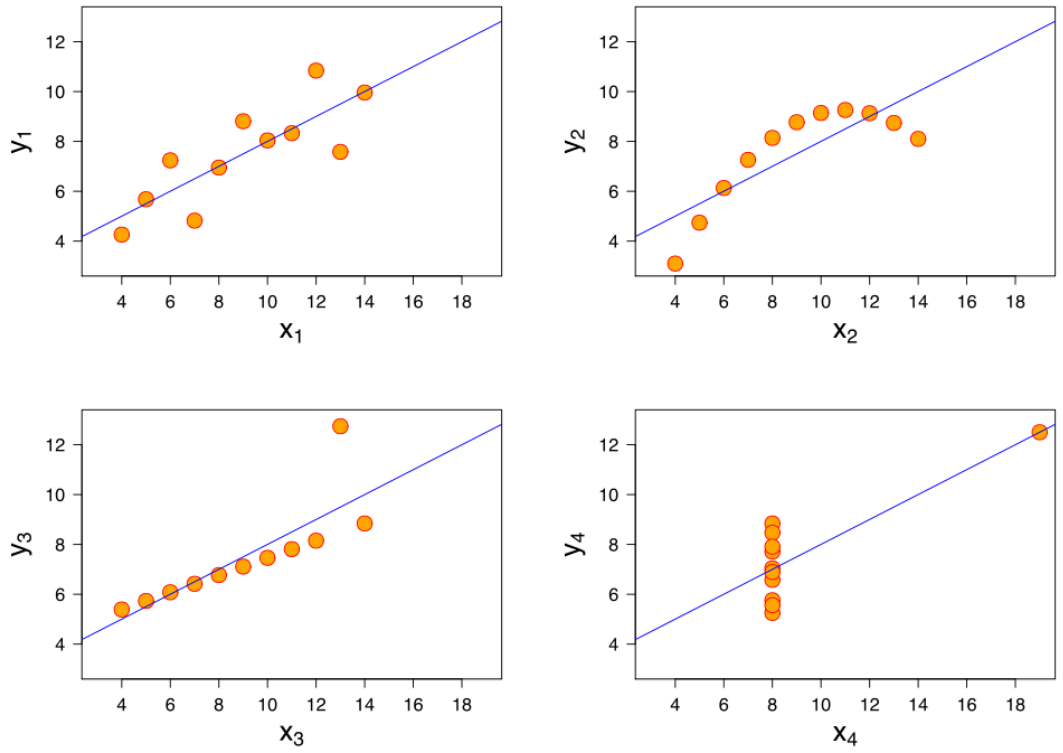
2. Does a correlation of r = 0 mean there is no association between two variables?



r = -0.08

Review of correlation cautions

3. Is the correlation coefficient (r) resistant?



Anscombe's quartet
r = 0.81

Q: What should we always do first when analyzing data?

A: Plot it!

Where do samples/data come from?

In past classes I have had students try out sampling by counting 100 sprinkles…

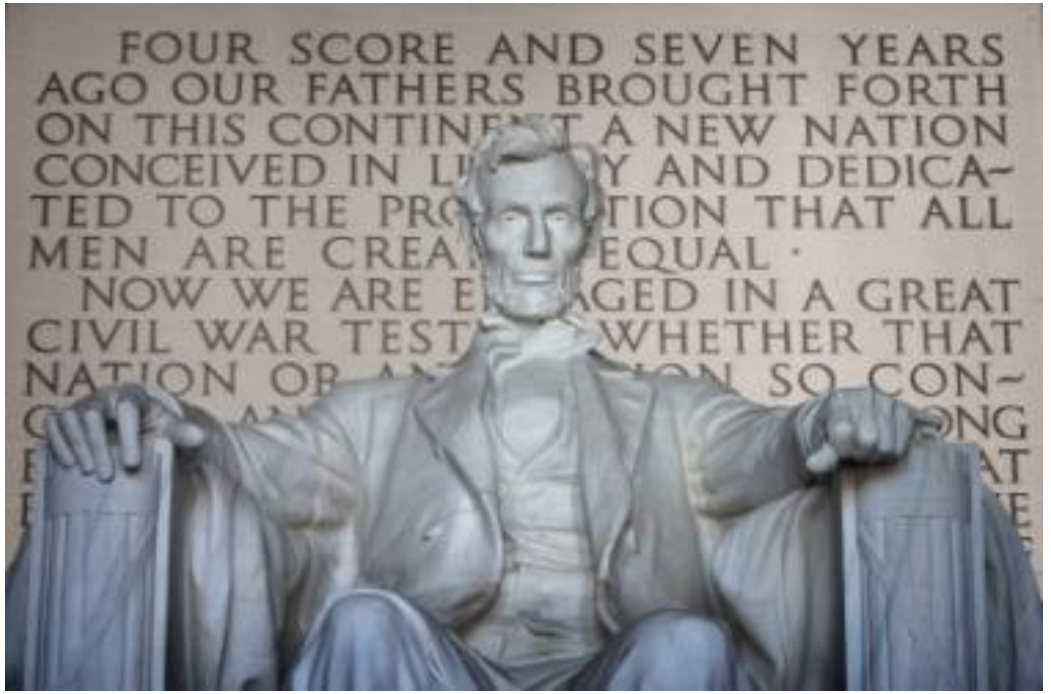
1	orange
2	red
3	green
4	white
5	white
6	white
7	white
8	white
9	red

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

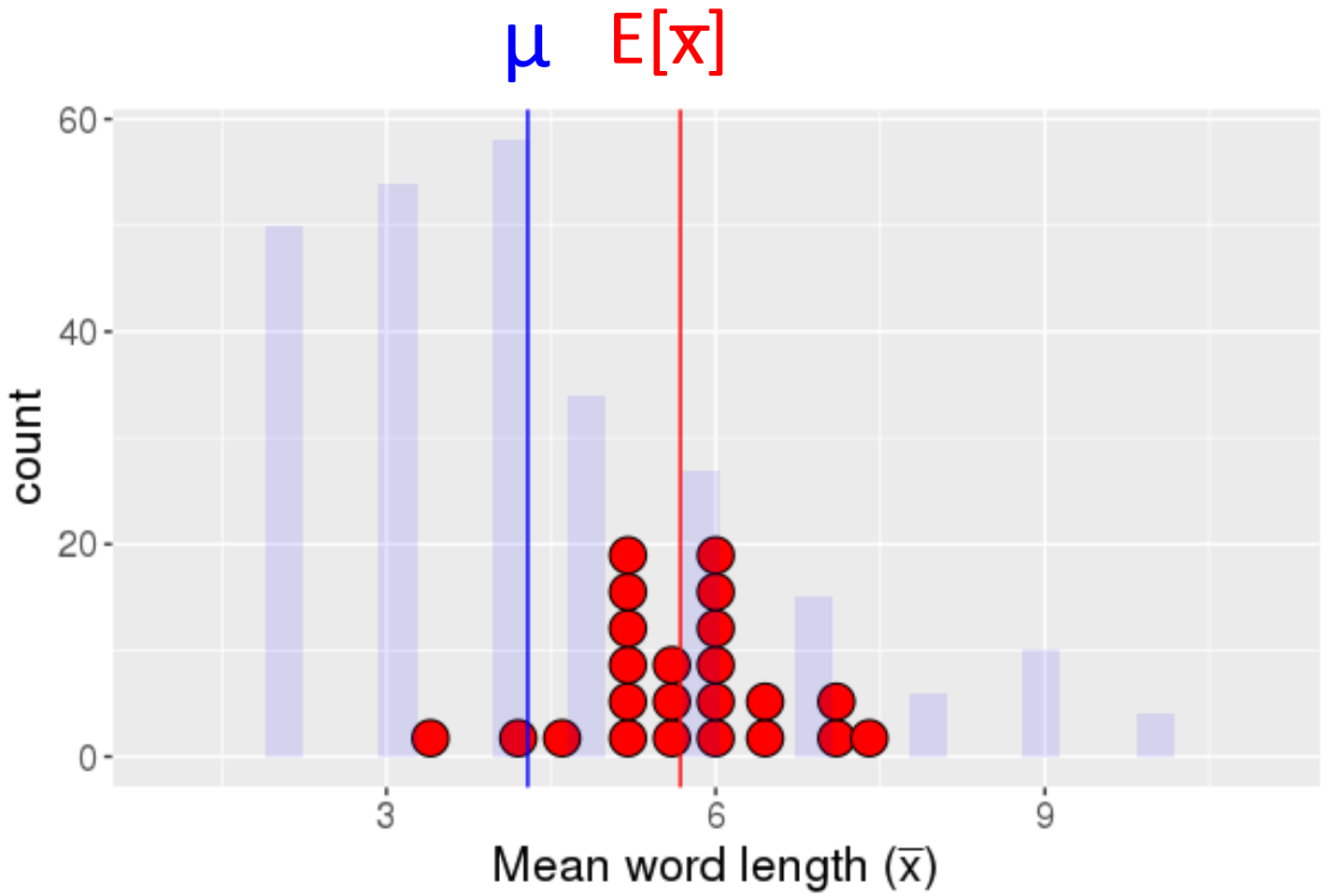
Let’s try some sampling ourselves…

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

Report the mean of the 10 words

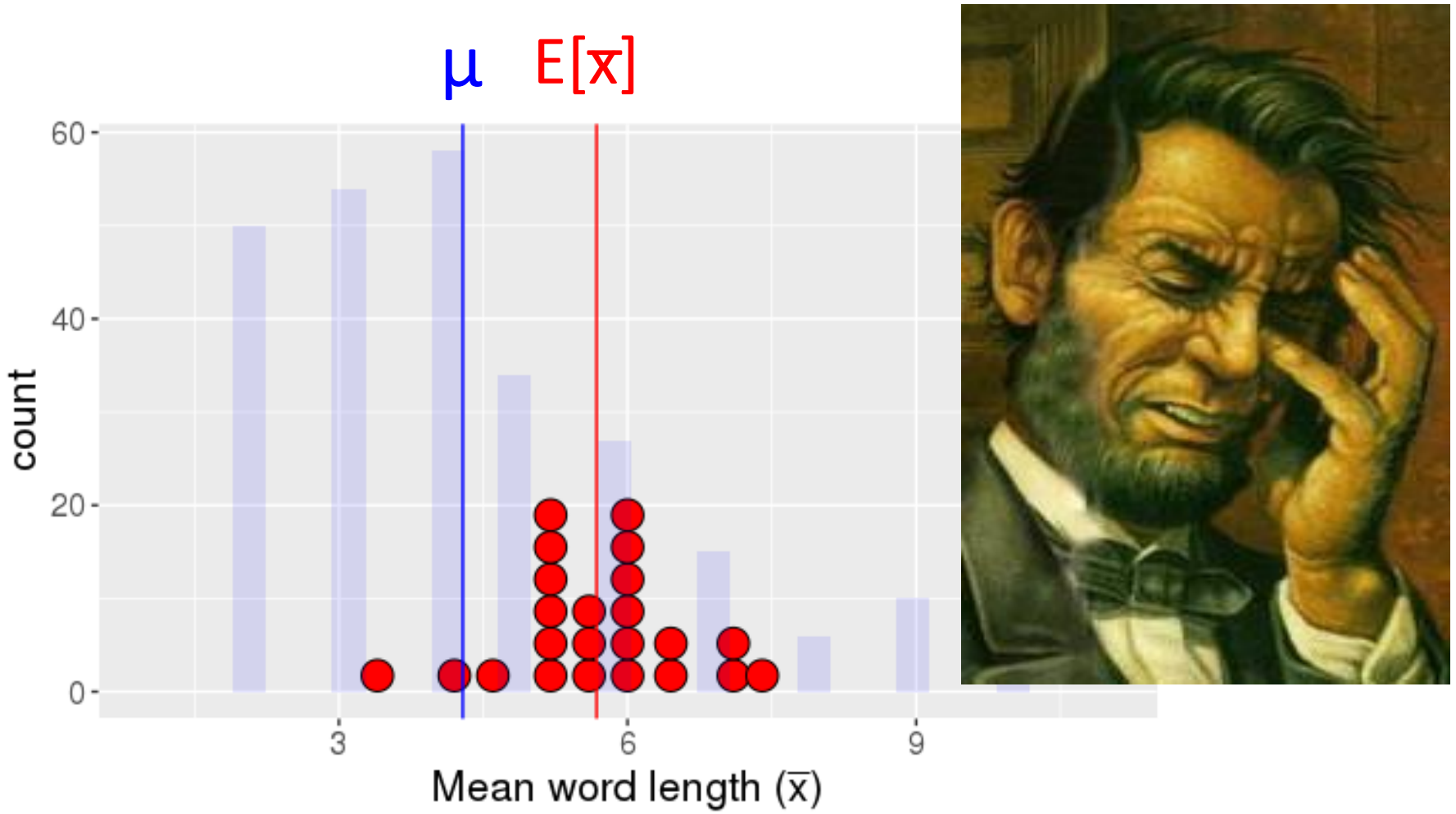


Gettysburg address, mean word length



Gettysburg address, mean word length

Observations?

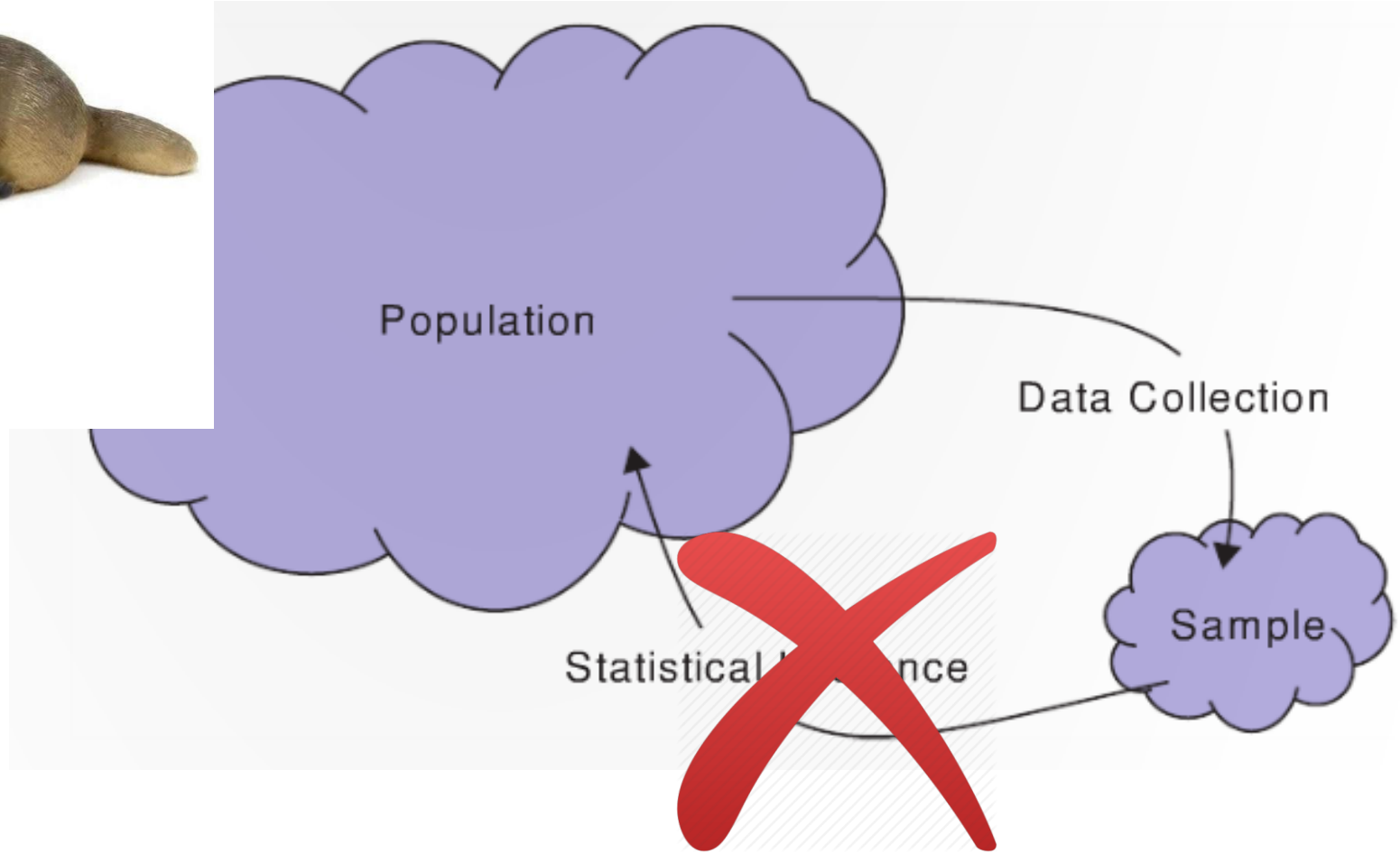


Other types of bias

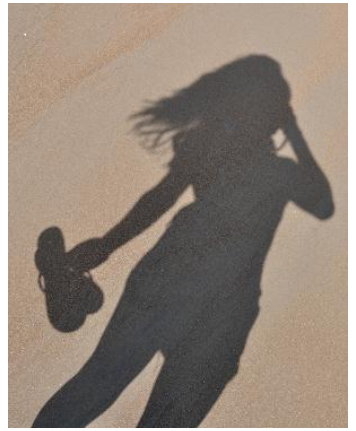
Bias exists when the method of collecting the data causes the sample to inaccurately reflect the population

Statistical bias

μ



$\bar{x}$



Dewey Defeats Truman? (1948)

The paper 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?

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

Alumni Survey Results

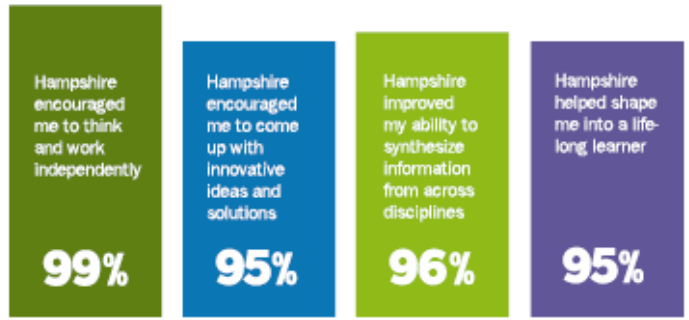
Hampshire College

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.

Not a problem as long as the sample is representative of the population

Example: Want to know proportion of people left-handed in the US. Randomly sample Hampshire students might be good enough.

Need to think carefully to avoid bias!

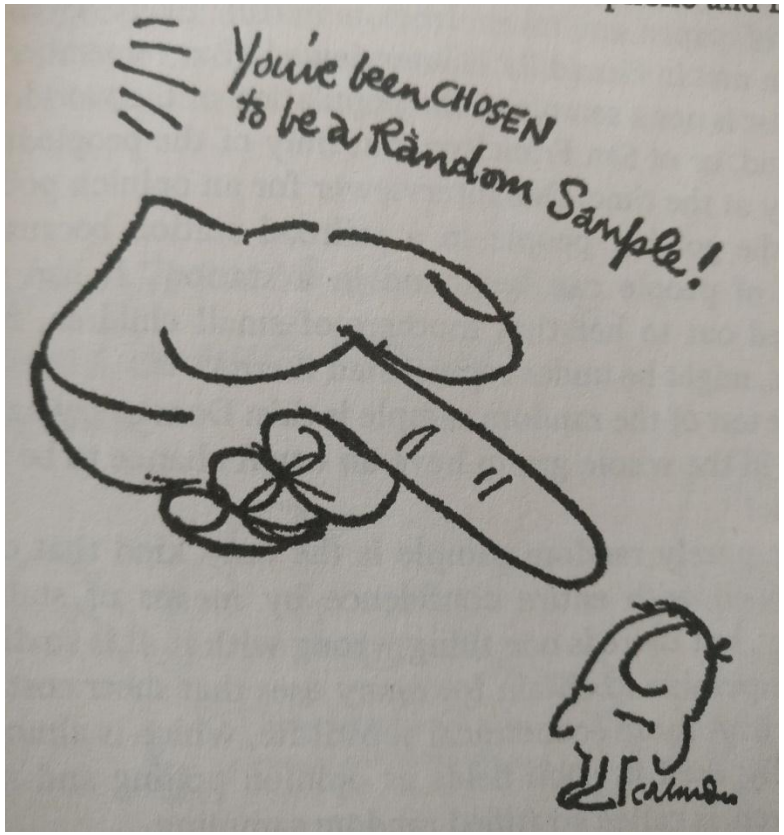
As mentioned last class, statistics requires thought!

Use your own reasoning:

Does the sample reflect the population of interest?

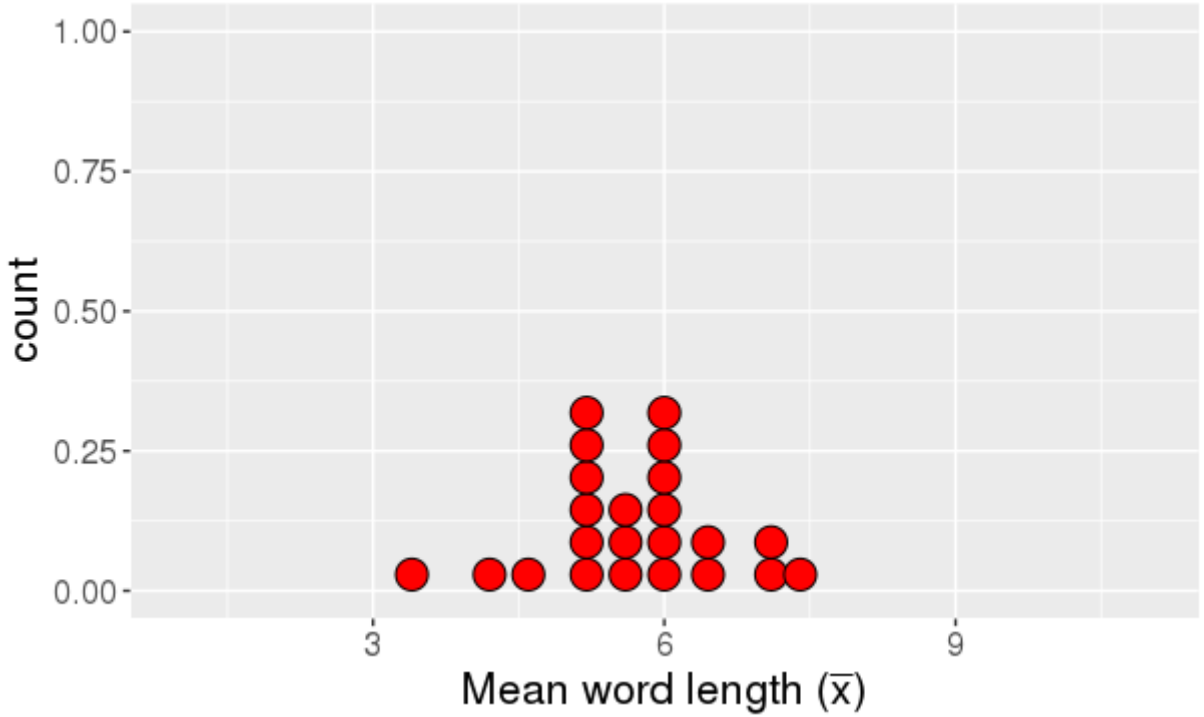
Be your own worst critic!

Questions about statistical bias?



For our distribution of Gettysburg word lengths…

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



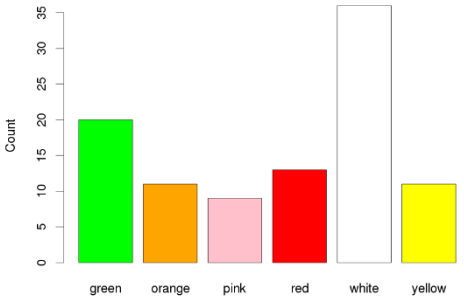
A: The mean length of 10 words ($\bar{x}$)

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

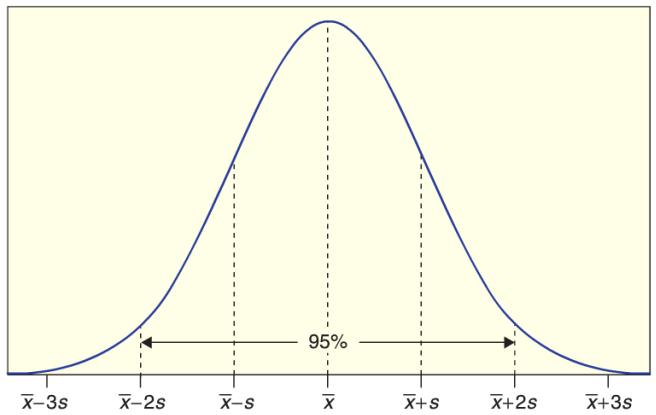
Sampling distribution

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

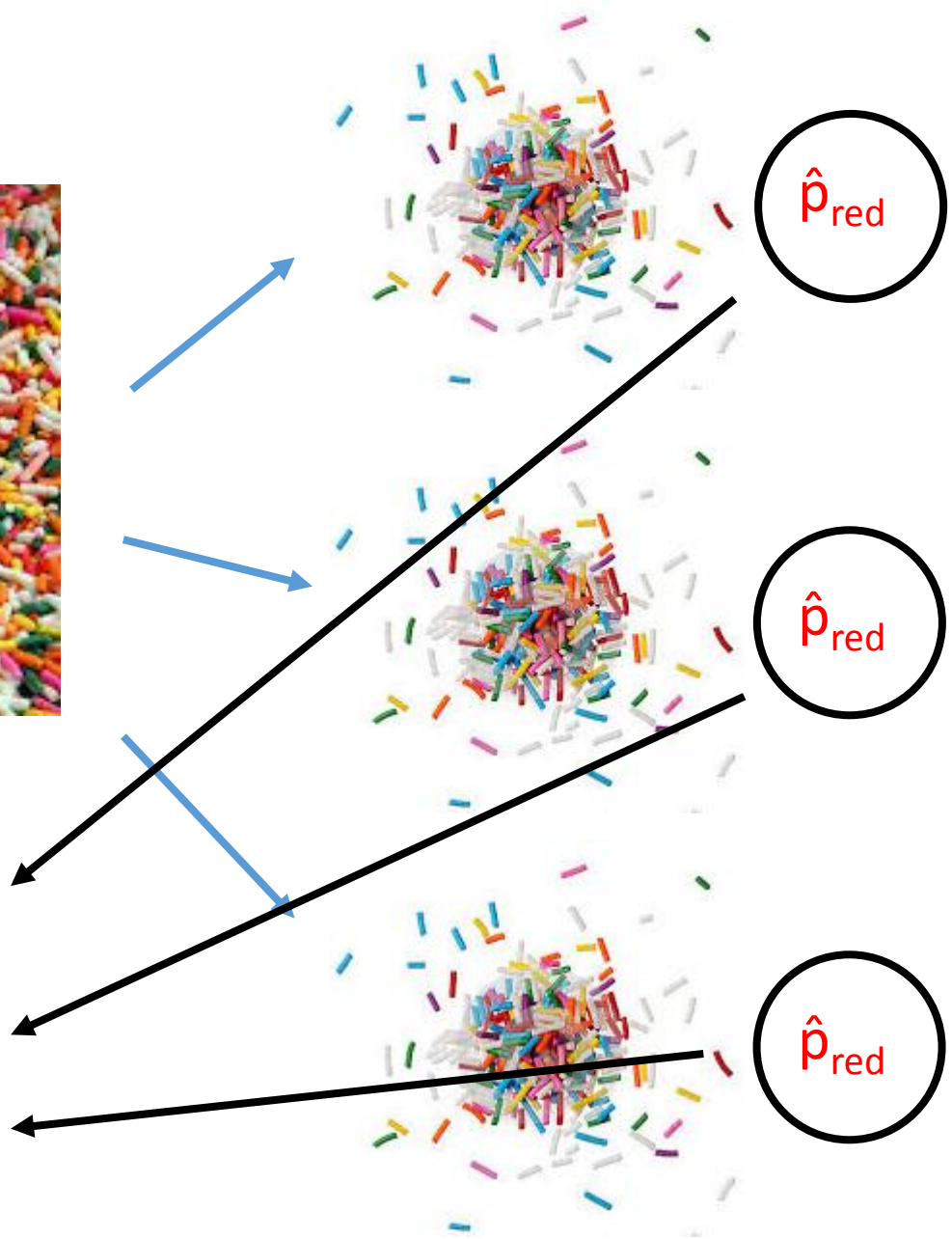
A sampling distribution shows us how the sample statistic varies from sample to sample



π_{red}



Sampling distribution!



Let’s create a sampling distribution in R

Get the Gettysburg population data

```
> load("/home/shared/intro_stats/cs206_data/gettysburg.Rda")
> word_lengths <- gettysburg$num_letters
```

We can use the sample(data_vec, n) to get a sample of length n

```
> curr_sample <- sample(word_lengths, 10)
```

Let’s store the mean of these 10 words in a vector called sample_means

```
> sample_means <- NULL
>  sample_means[1] <- mean(curr_sample)
```

Repeat this many times to get an approximation of the sampling distribution and plot them as a histogram…

Let’s create a sampling distribution in R

```
> curr_sample <- sample(word_lengths, 10)
> sample_means[2] <- mean(curr_sample)

> curr_sample <- sample(word_lengths, 10)
> sample_means[3] <- mean(curr_sample)

 …

> curr_sample <- sample(word_lengths, 10)
> sample_means[100] <- mean(curr_sample)

# create a histogram to see the shape of the sampling distribution
> hist(sample_means)
```

Let’s create a sampling distribution in R

Writing the same code again and again is tedious, let’s use a for loop

```
for (i in 1:1000) {
        print(i)
}
```

Let’s create a sampling distribution in R

Writing the same code again and again is tedious, let’s use a for loop

```
the_squares <- NULL
for (i in 1:1000) {
        the_squares[i] <- i^2
}

the_squares
```

Let’s create a sampling distribution in R

Writing the same code again and again is tedious, let’s use a for loop

```
sample_means <- NULL

for (i in 1:1000) {

        curr_sample <- sample(word_lengths, 10)
        sample_means[i] <-  mean(curr_sample)

}

hist(sample_means)
```

Worksheet 5

Answer a few questions about representative samples, and get practice using for loops:

```
> source("/home/shared/intro_stats/cs206_functions.R")
> get_worksheet(5)
```

Due at 11:59pm on Sunday October 14th