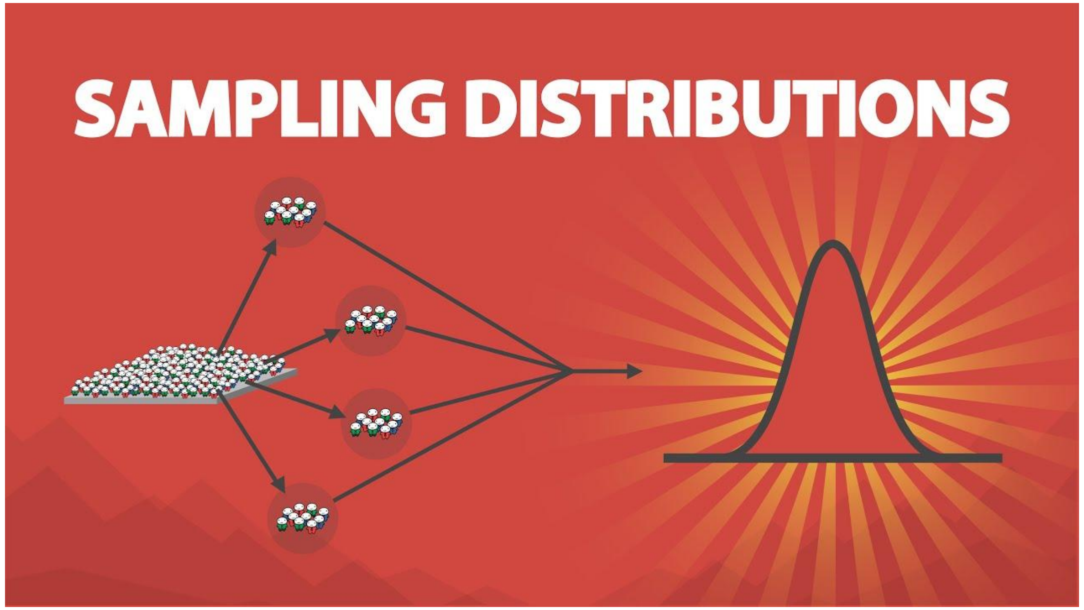
Sampling distributions, standard errors, and the mini-exam



Overview

Review and continuation of sampling and bias

Sampling distributions and the standard error

If there is time:

- Exploring sampling distributions in R

The mini-exam

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

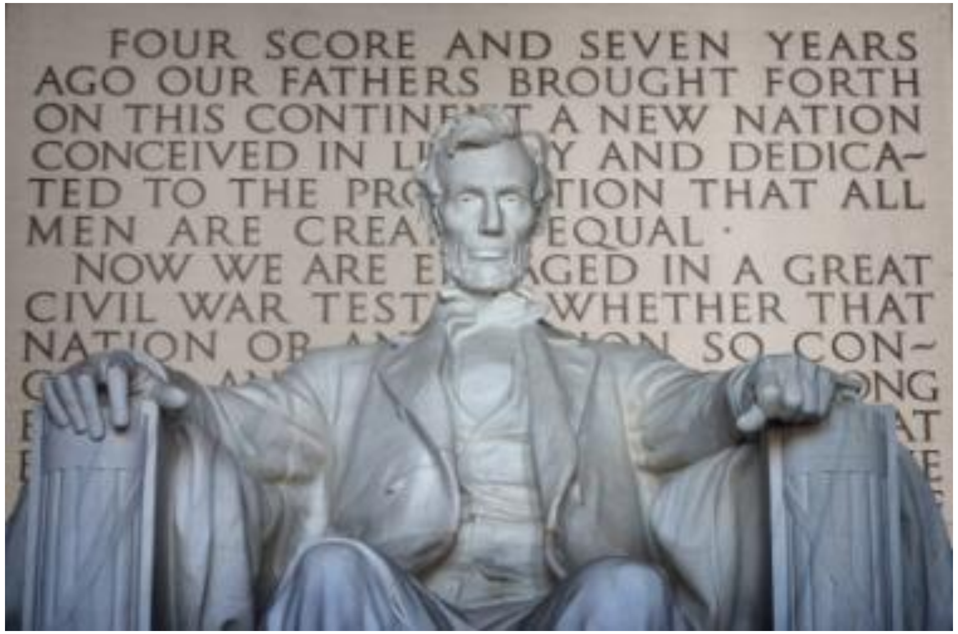
- Time: Sunday 1-2pm
- Location: Bass L01A

Also, keep attending the practice sessions for more practice!

Review: sampling and sampling distributions

Review: sampling

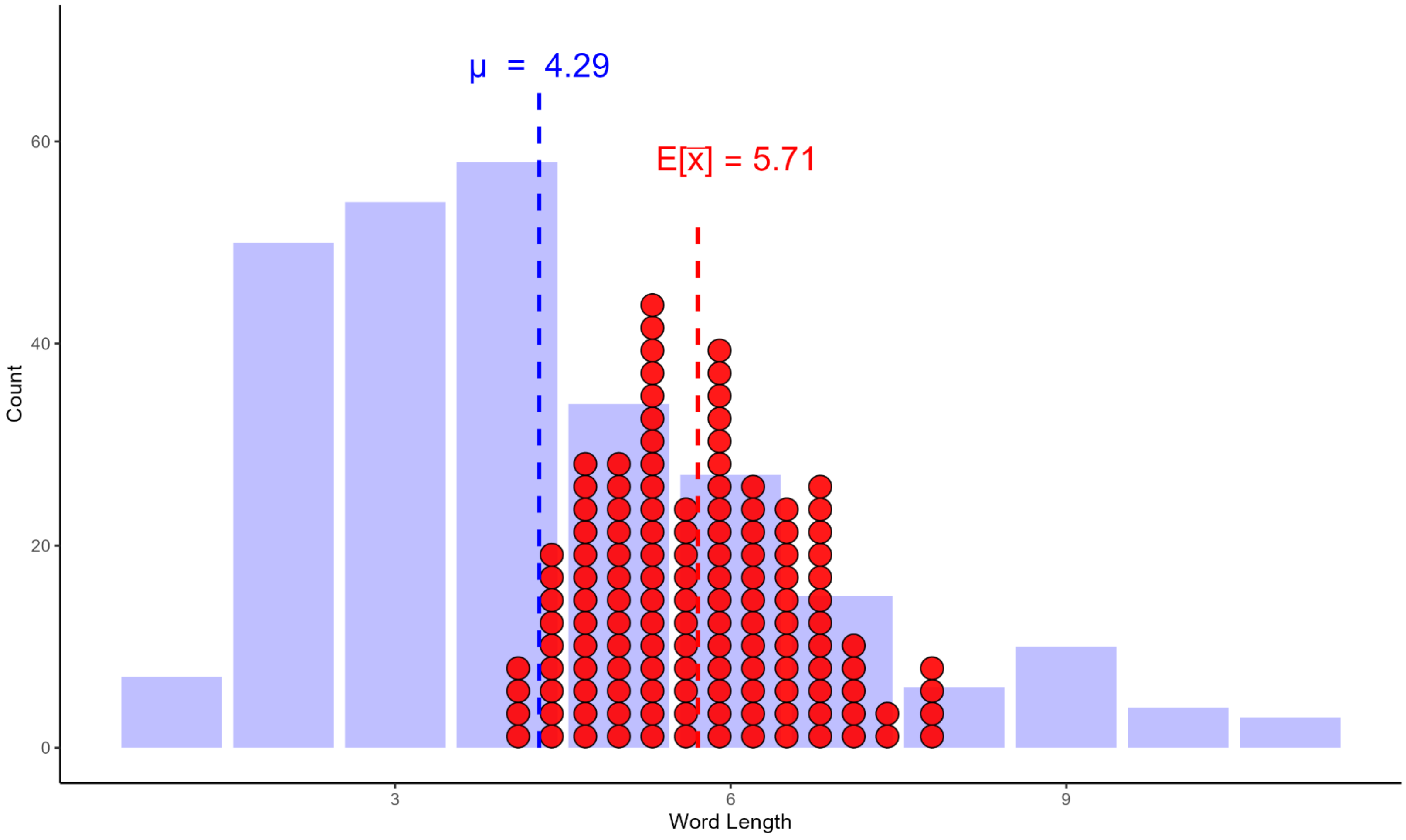
1	orange
2	red
3	green
4	white
5	white
6	white
7	white
8	white
9	red



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

A: ***n***

Bias and the Gettysburg address word length distribution

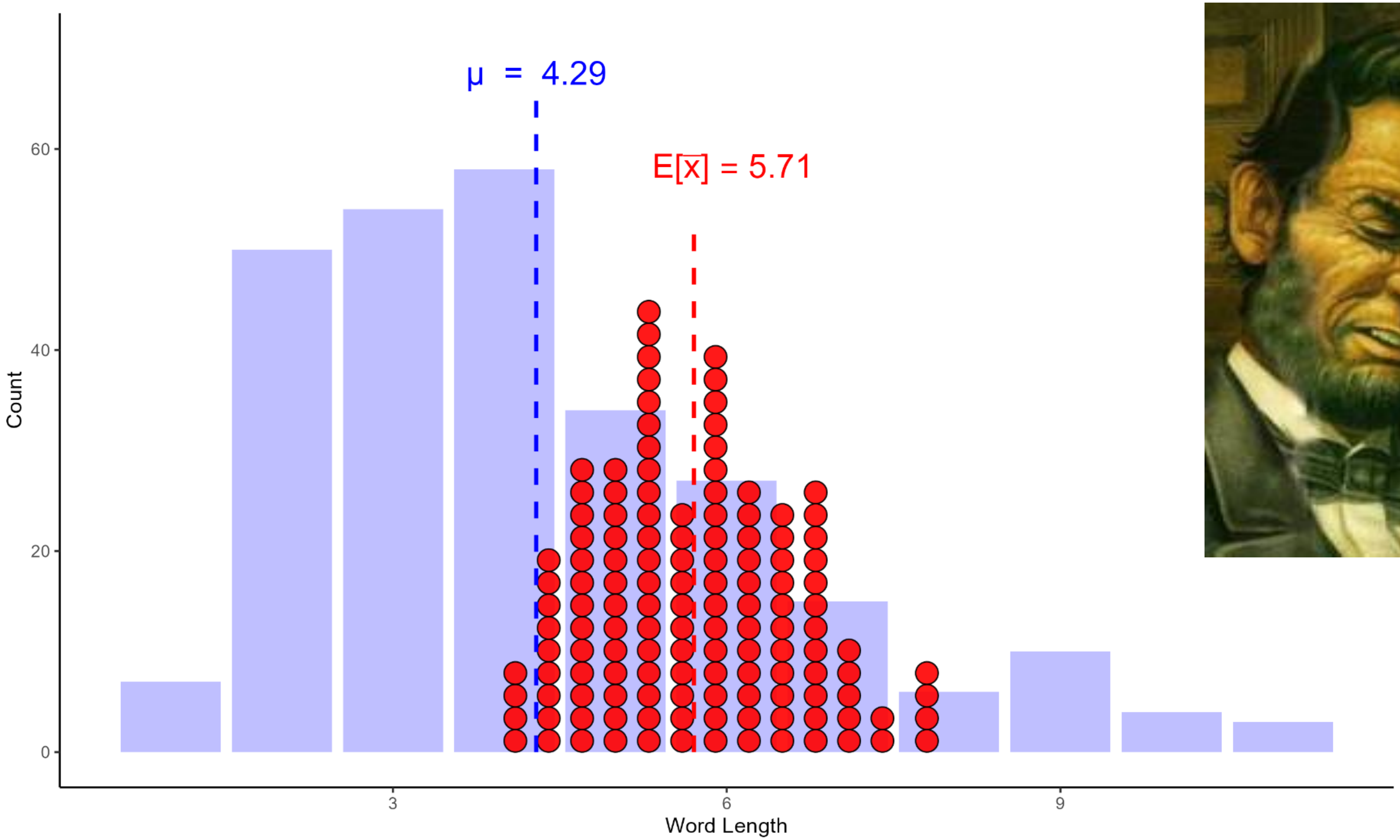


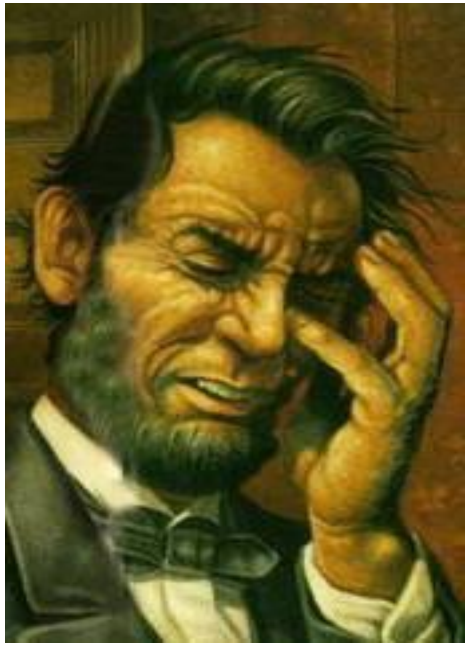
Bias and the Gettysburg address word length distribution

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

Here:

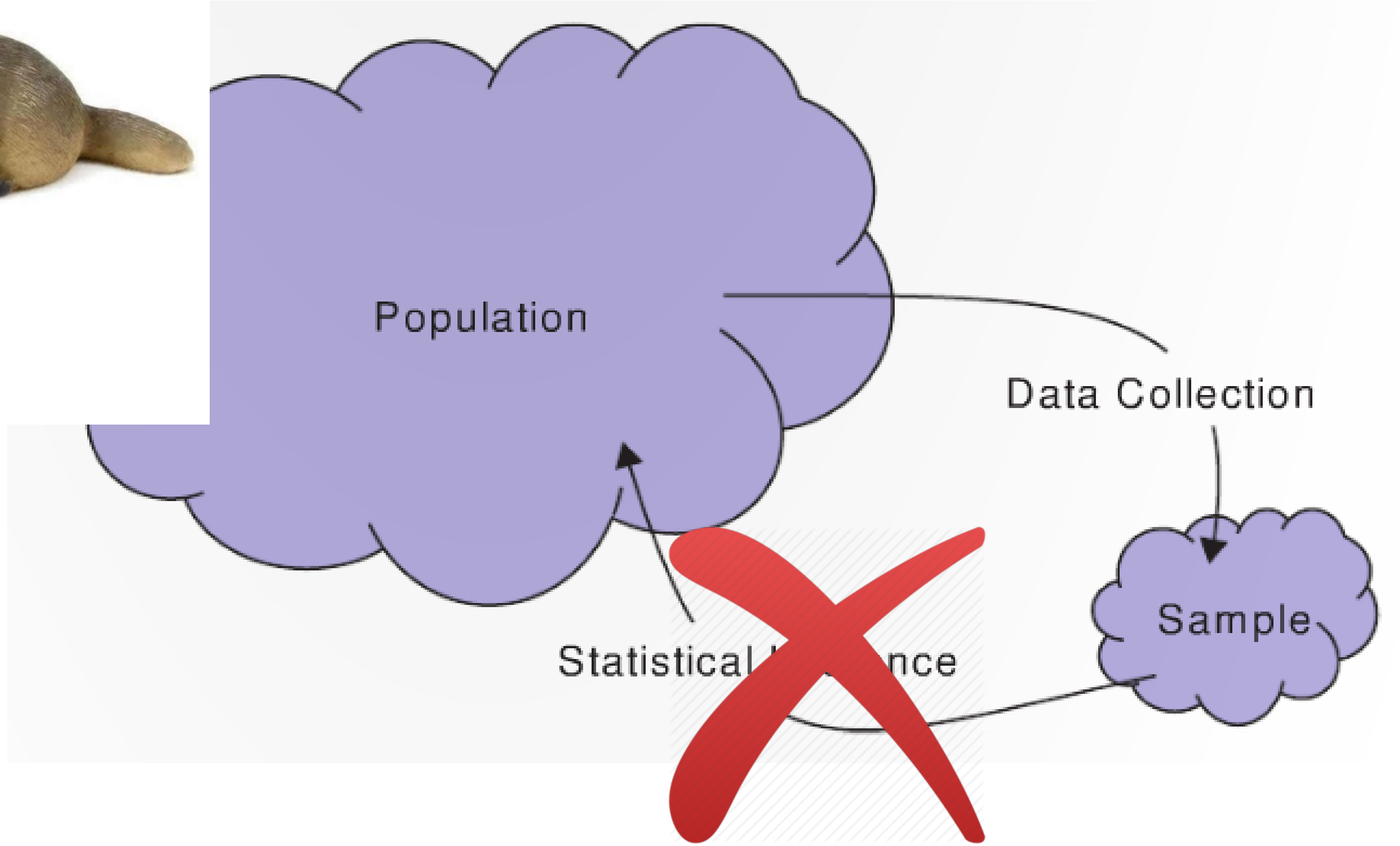
$E_s[\overline{x}] \neq \mu$





Statistical bias

μ



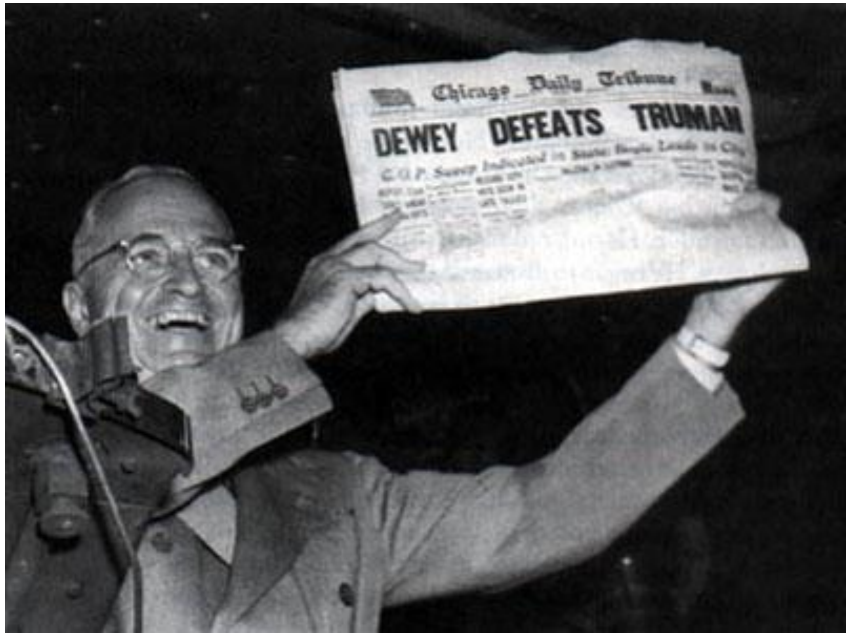
$\bar{x}$

Basic questions for sampling

What is the population?

What is the sample?

Do they differ in a meaningful way?



Bias or no bias?

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 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%

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 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!

Q: How can we prevent sampling bias?

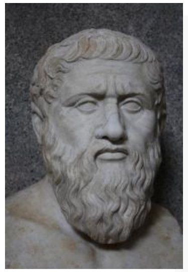
A: To prevent bias, use a **simple random sample**

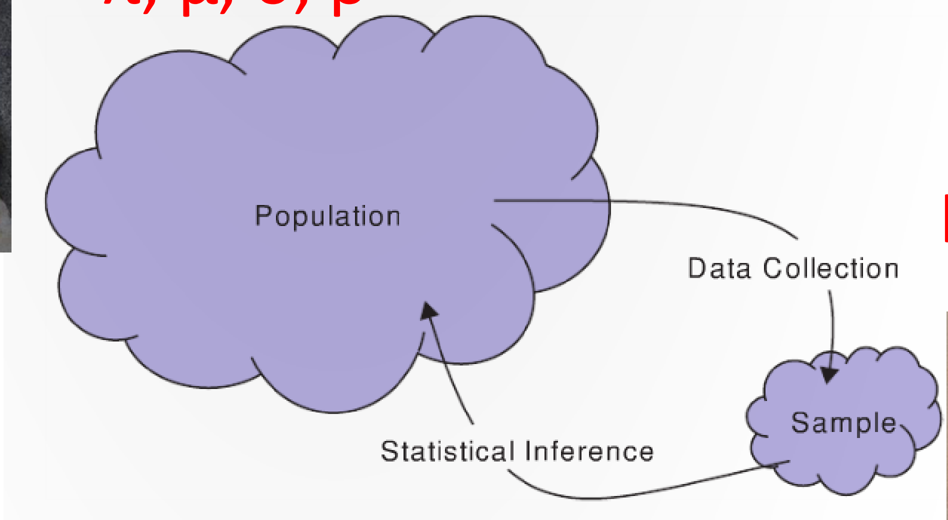
- where each member in the population is equally likely to be in the sample

This allows for generalizations to the population!

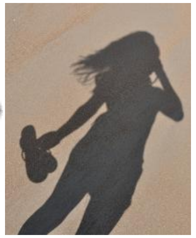
Soup analogy!

π, μ, σ, ρ





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



Q: How do we select a random sample?

Mechanically:

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

Use a computer program

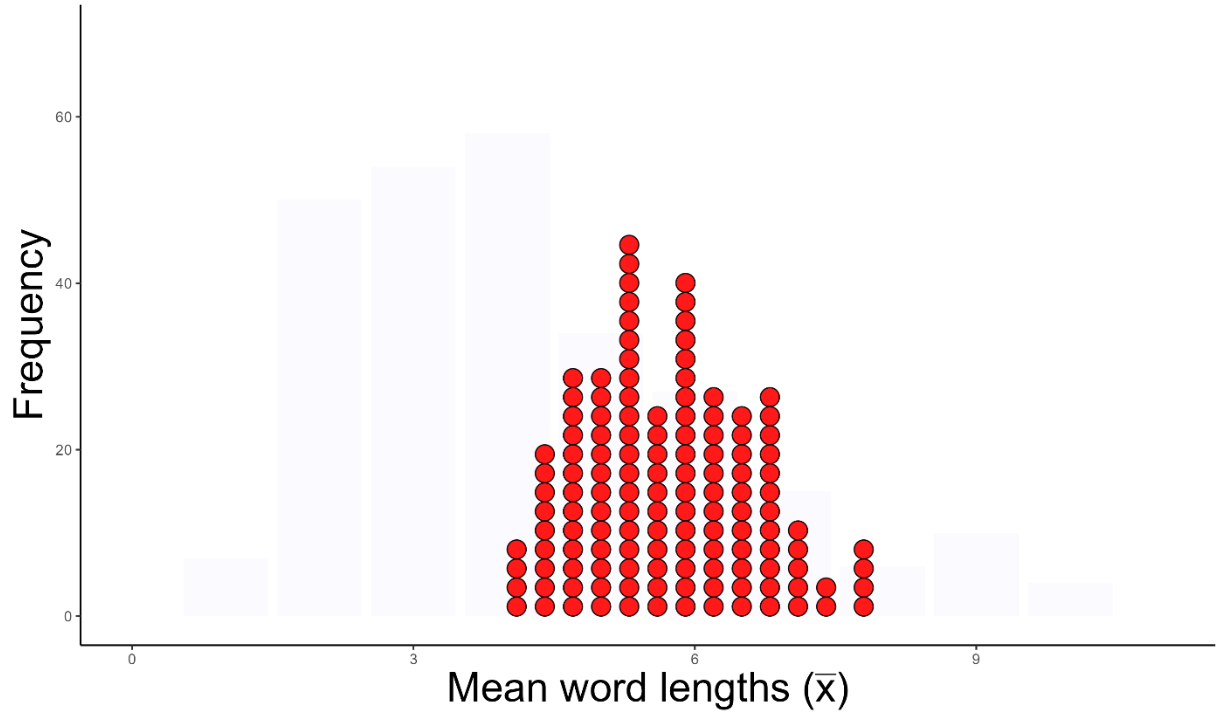
From now on we are going to assume no bias!

Our statistic values, on average, reflect the parameters

Sampling distributions

Recall for our distribution of Gettysburg word lengths…

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



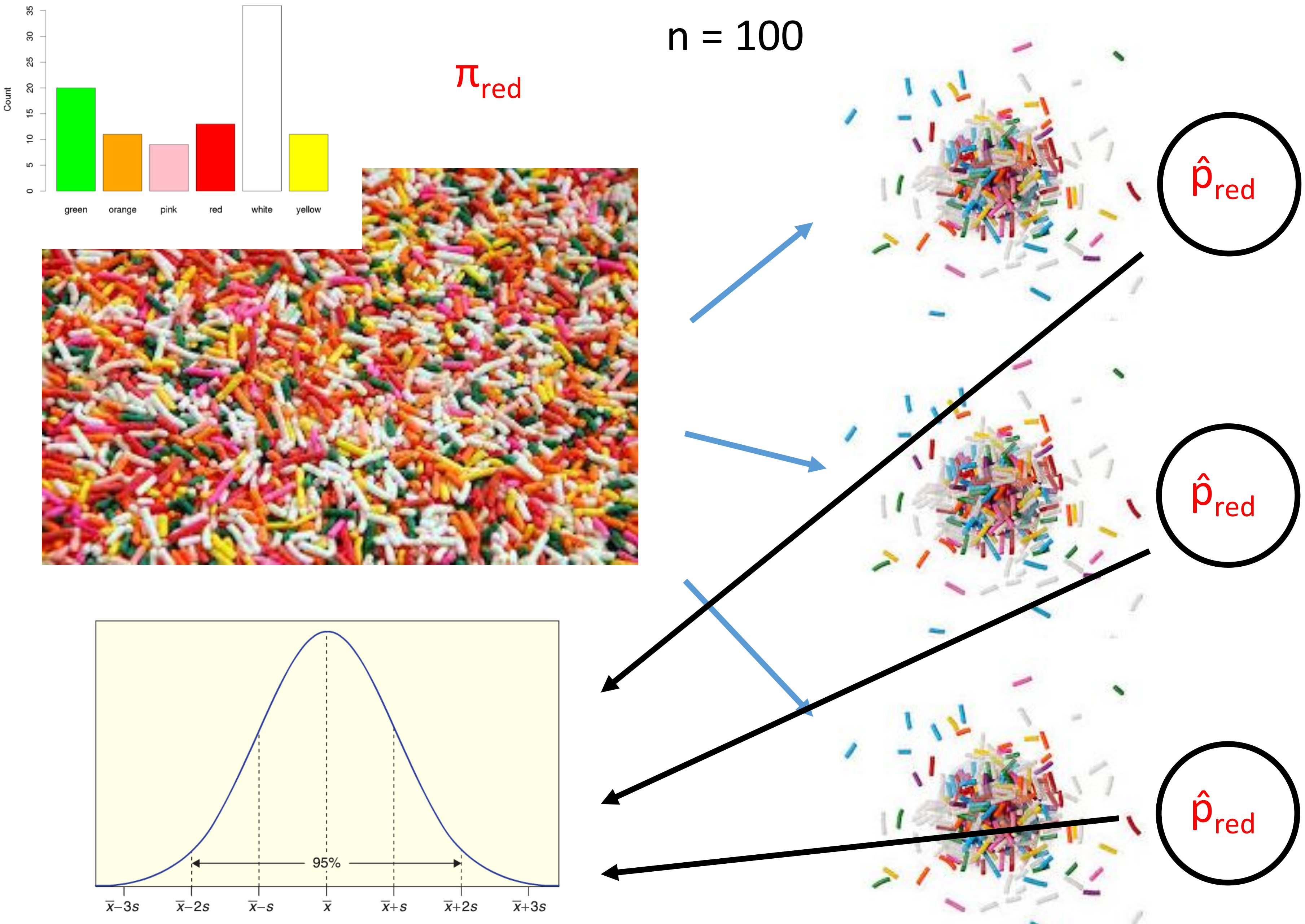
A: The mean length of 10 words ($\bar{x}$)

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

Sampling distribution

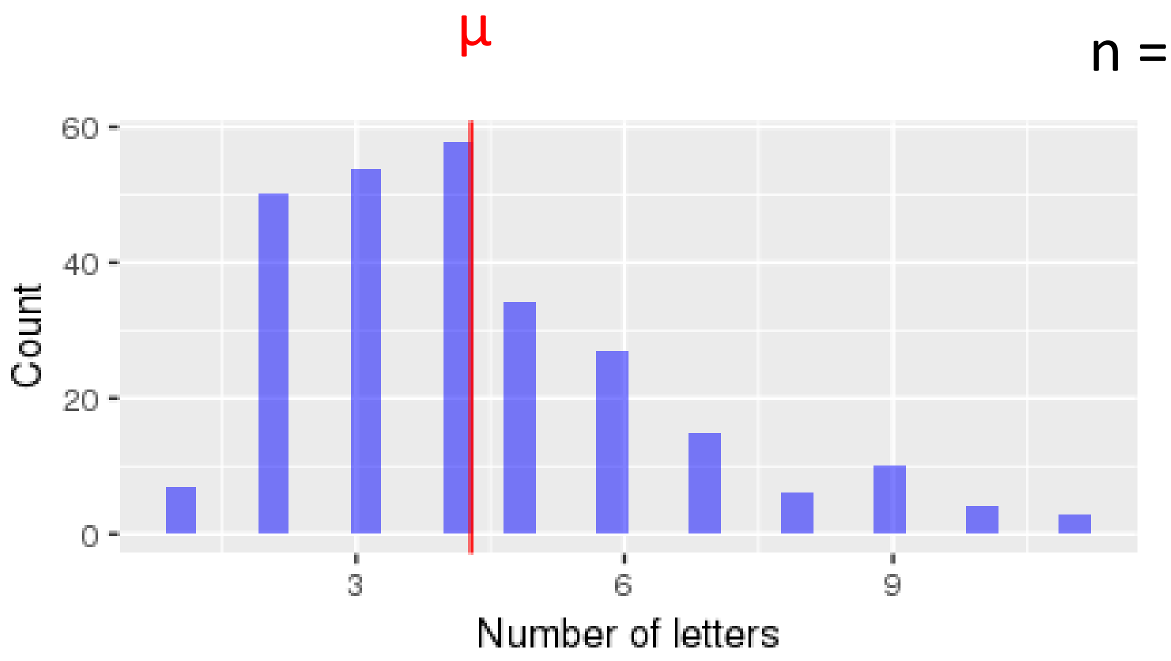
A **sampling distribution** is the distribution of <u>sample statistics</u> computed from different samples of the same size (n) from the same population

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



Sampling distribution!

Gettysburg address word length sampling distribution



n = 10

10, 3, 3, 3, 4,
3, 2, 6, 10, 5

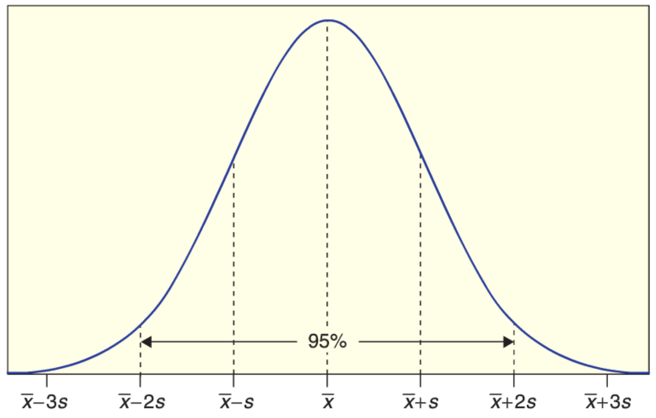
$\overline{x} = 5$

2, 6, 2, 6, 6,
2, 5, 3, 2, 9

$\overline{x} = 4.3$

3, 9, 3, 4, 4,
3, 6, 6, 2, 2

$\overline{x} = 4.2$



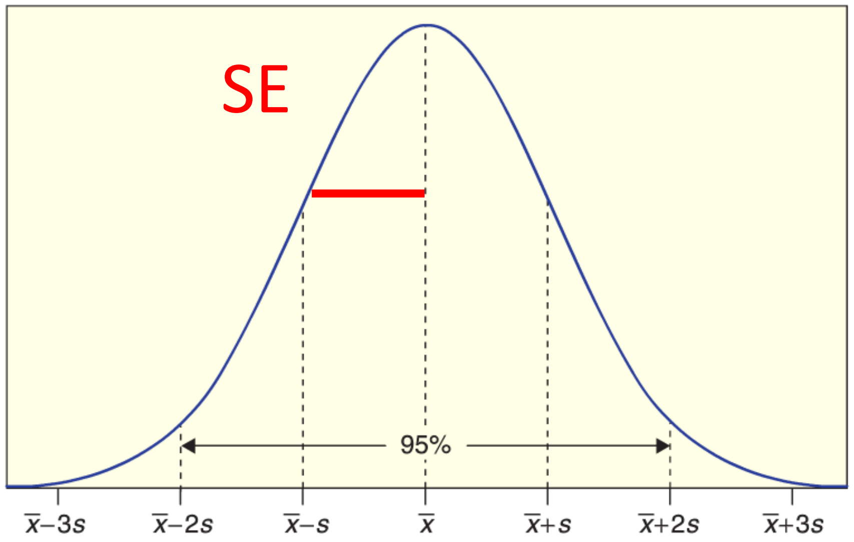
Sampling distribution!

[Gettysburg sampling distribution app]

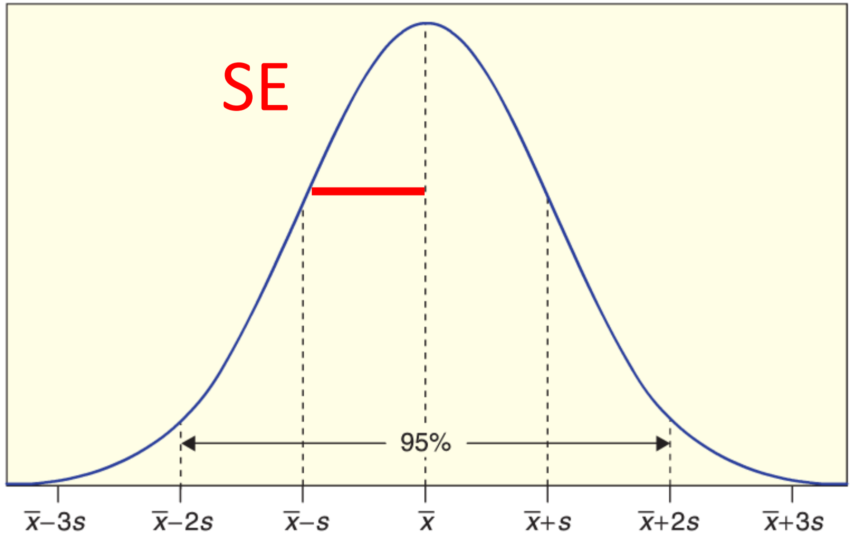
The standard error

The **standard error** of a statistic, denoted SE, is the standard deviation of the sample statistic

- i.e., SE is the standard deviation of the *sampling distribution*



What does the size of a standard error tell us?



Q: If we have a large SE, would we believe a given statistic is a good estimate for the parameter?

- E.g., would we believe a particular $\bar{x}$ is a good estimate for μ?

A: A large SE means our statistic (point estimate) could be far from the parameter

- E.g., $\bar{x}$ could be far from μ

Sampling distributions in R!

Let’s create a sampling distribution in R

Load the SDS1000 library to make all SDS1000 functions available

```
library(SDS1000)
```

Get the Gettysburg population data

```
load("gettysburg.Rda")
word_lengths <- gettysburg$num_letters     # lengths of the 268 words
```

Let's create a sampling distribution in R

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

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

Q: How can we get $\bar{x}$ from this sample in R?

```
mean(curr_sample)
```

Q: How could we get a full sampling distribution?

- A: Repeat this many times to get an approximation of the sampling distribution
- If we store the $\bar{x}$'s in a vector, we can then plot the sampling distribution as a histogram

The do_it() function

The do_it() function (from the SDS1000 package) repeats a piece of code many times

- It returns a vector with the values created each time the code is repeated

```
do_it(100)  *  {

        2 + 3

}
```

Let's create a sampling distribution in R

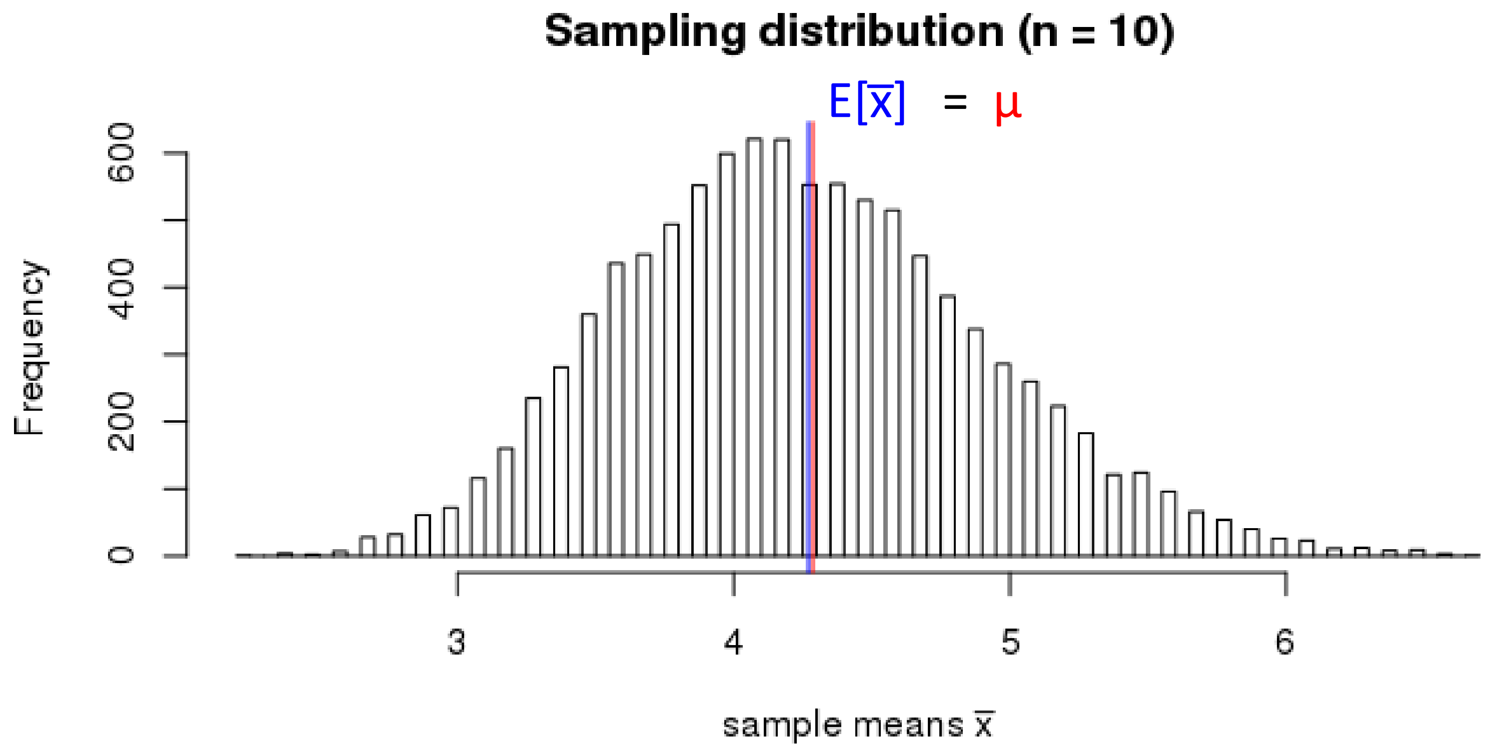
```
sampling_dist <- do_it(10000)  *  {

        curr_sample <- sample(word_lengths, 10)
        mean(curr_sample)

}

hist(sampling_dist)
```

Sampling distribution in R



mean(sampling_dist)
mean(word_lengths)   # these are the same, so no bias

Changing the sample size n

What happens to the sampling distribution as we change ***n***?

- Experiment for n = 1, 5, 10, 20

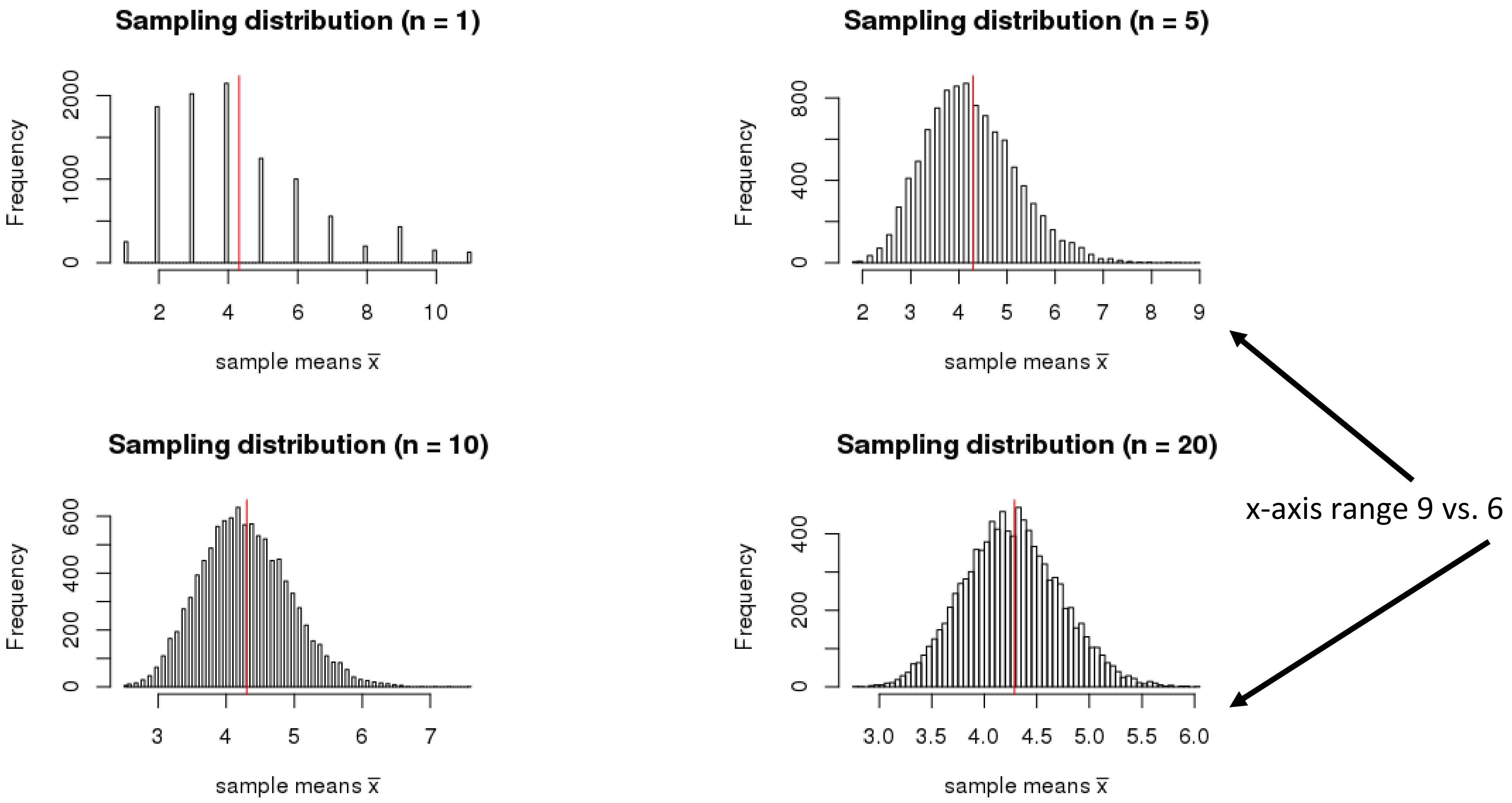
```
sampling_dist <- do_it(10000)  *  {

        curr_sample <- sample(word_lengths, 20)
        mean(curr_sample)

}

hist(sample_means, breaks = 100)
```

Gettysburg sampling distribution app



As the sample size n increases

1. The sampling distribution becomes more like a normal distribution
2. The sampling distribution points ($\bar{x}$'s) become more concentrated around the mean $E[\bar{x}] = \mu$