

Confidence intervals, sampling
distributions, and standard errors

Overview

Questions about worksheet 5?

Point estimates and confidence intervals

Review: sampling bias and sampling distributions

More on sampling distributions and the Standard Error

Questions about worksheet 5?

How does everyone feel about for loops?

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

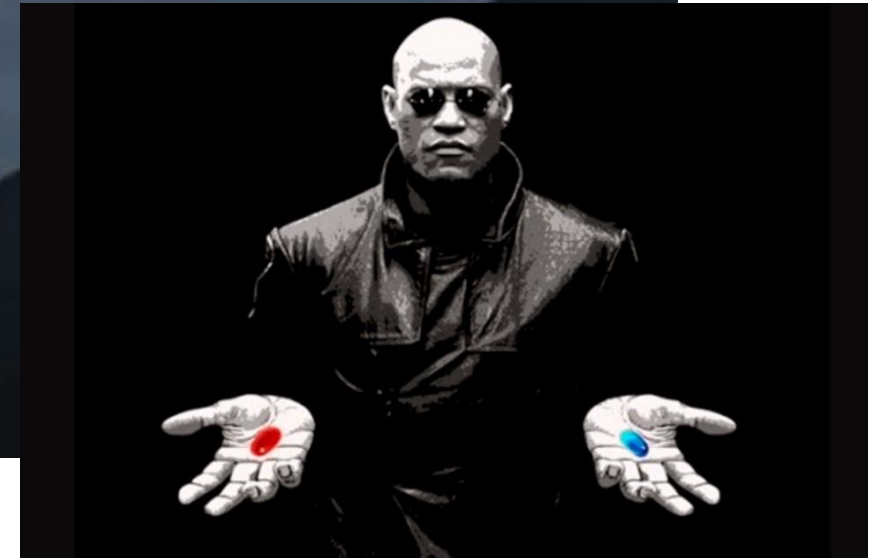
Questions about worksheet 4?

How does everyone feel about for loops?

```
my_vec <- NULL
for (i in 1:1000) {
  my_vec[i] <- i * 3
}
```

You will get more practice soon!

The big picture

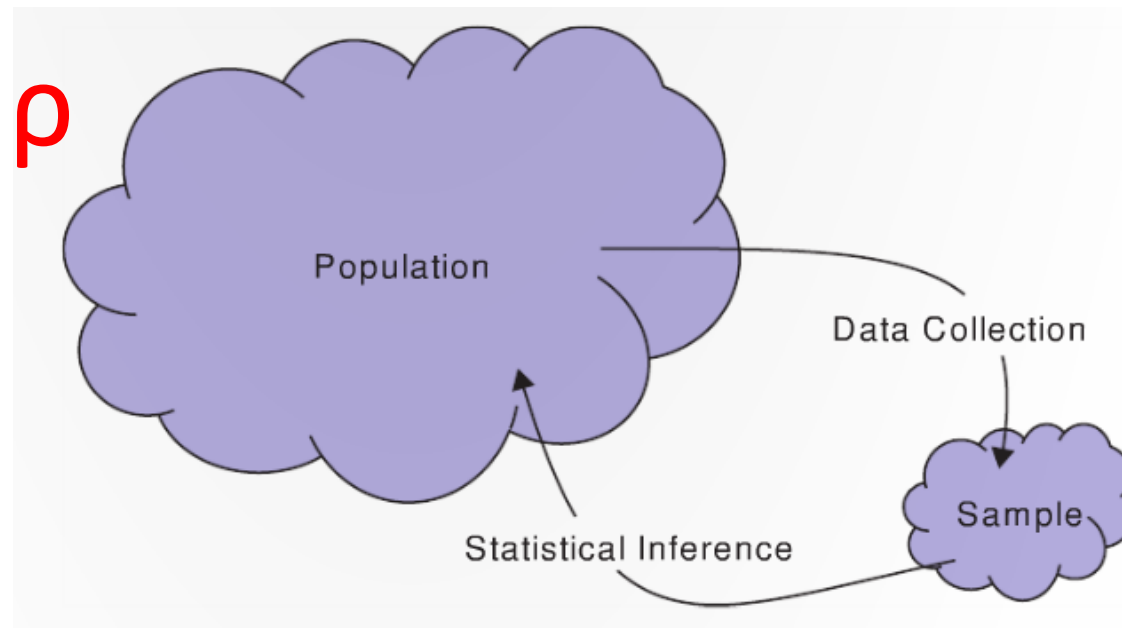
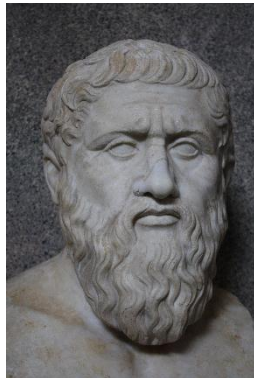


Back to the big picture: Inference

Statistical inference is...?

the process of drawing conclusions about the entire population based on information in a sample

π, μ, σ, ρ



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



Point Estimate

We use the statistics from a sample as a **point estimate** for a population parameter

- $\bar{x}$ is a point estimate for...? μ

44% of American approve of Trump's job performance

- [Gallup poll from October 14th](#)

Q: What are π and $\hat{p}$ here?

Q: Is $\hat{p}$ a good estimate for π in this case?

A: We can't tell from the information given

Interval estimate based on a margin of error

An **interval estimate** give a range of plausible values for a population parameter.

One common form of an interval estimate is:

Point estimate $\pm$ margin of error

Where the **margin of error** is a number that reflects the precision of the sample statistic as a point estimate for this parameter

Example: Fox news poll

44% of American approve of Trump's job performance, plus or minus 3%

How do we interpret this?

Says that the population parameter (π) lies somewhere between 41% to 47%

i.e., if they sampled all voters the true population proportion (π) would be in this range

Confidence Intervals

A **confidence interval** is an interval computed by a method that will contain the *parameter* a specified percent of times

- i.e., if the estimation were repeated many times, the interval will have the parameter x% of the time

The **confidence level** is the percent of all intervals that contain the parameter

Think ring toss...

Parameter exists in the ideal world

We toss intervals at it

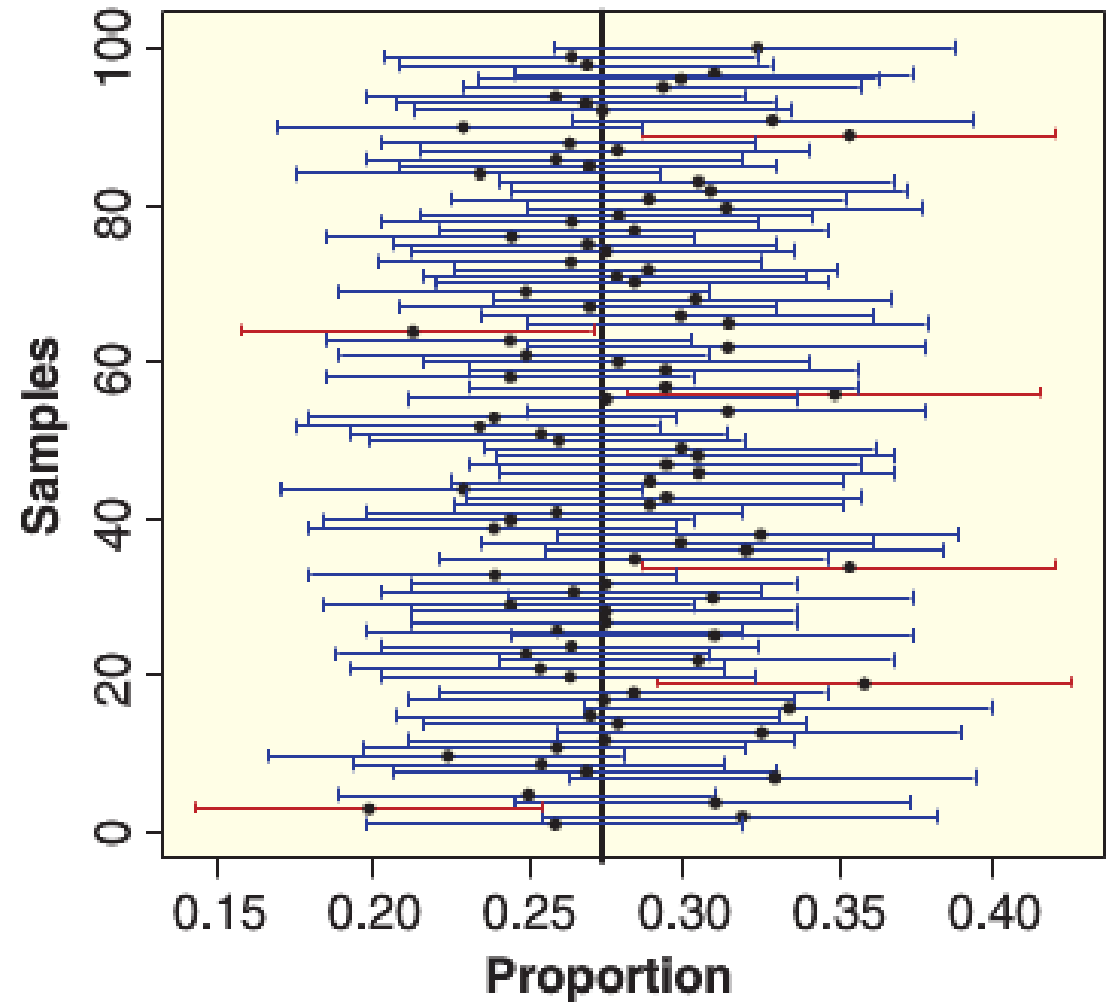
95% of those intervals capture the parameter



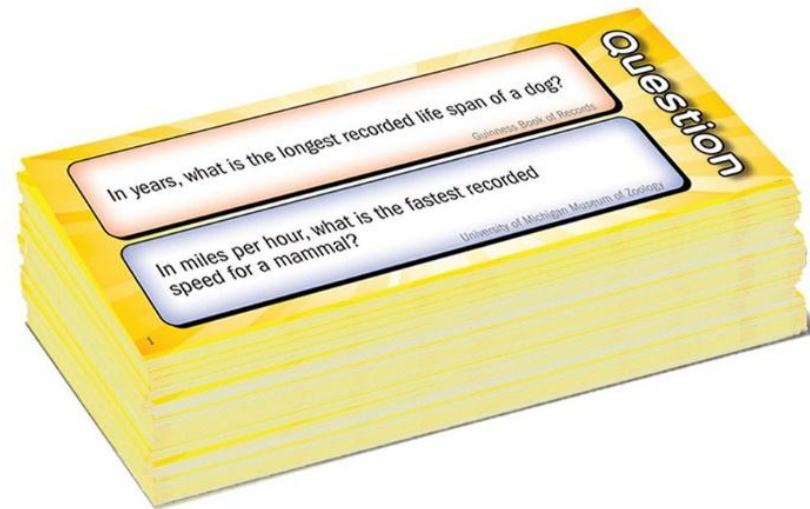
Confidence Intervals

For a **confidence level** of 95%...

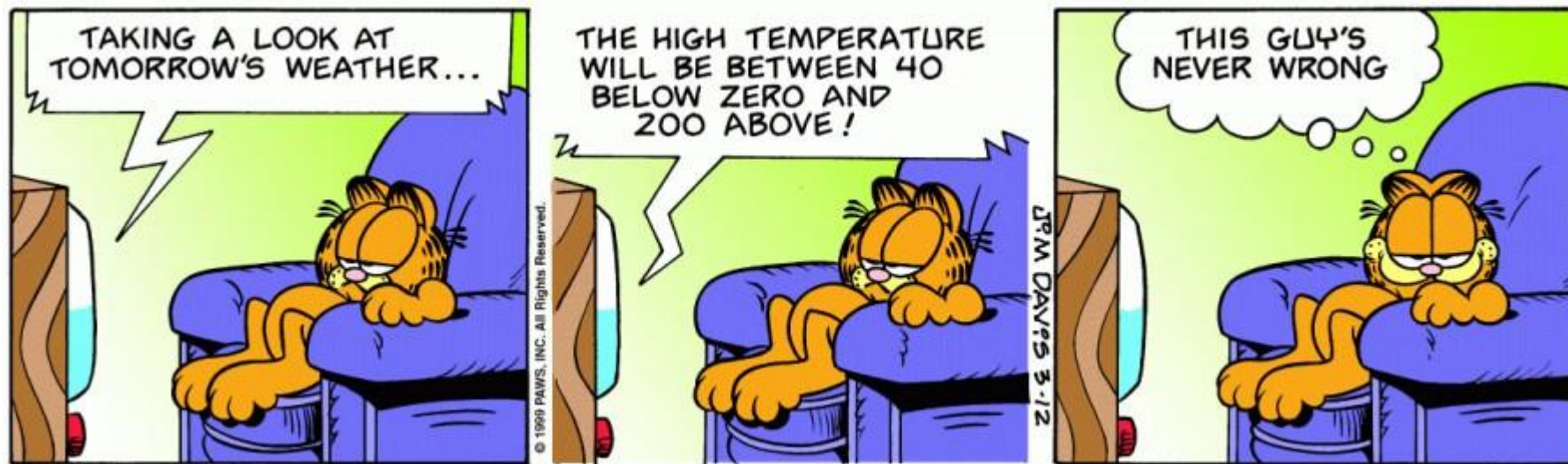
95% of the **confidence intervals** will have the parameter in them



Wits and Wagers...



100% confidence intervals



There is a tradeoff between the **confidence level** (percent of times we capture the parameter) and the **confidence interval size**

Note

For any given confidence interval we compute, we don't know whether it has really captured the parameter

But we do know that if we do this 100 times, 95 of these intervals will have the parameter in it

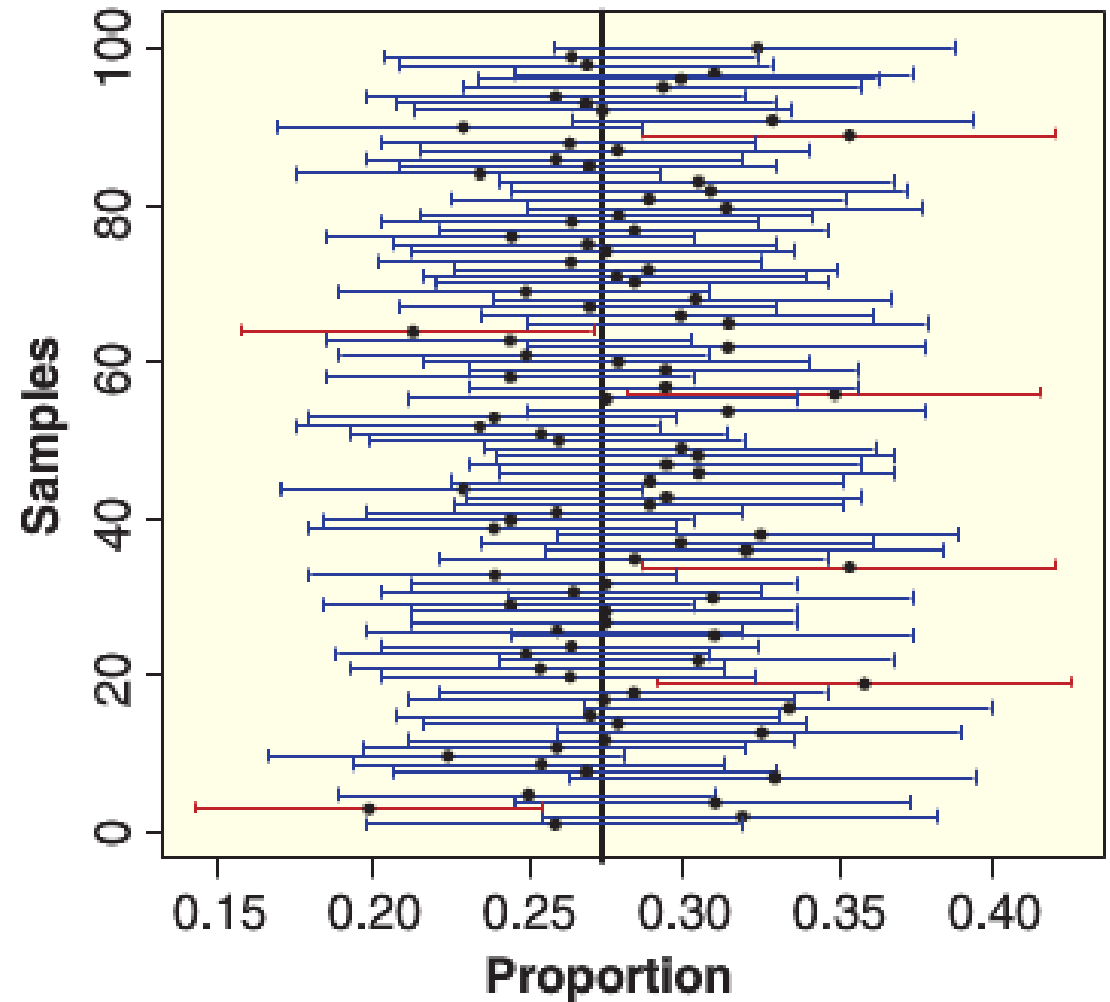
(for a 95% confidence interval)

Confidence Intervals

For a **confidence level** of 90%...

90% of the **confidence intervals** will have the parameter in them

Right!



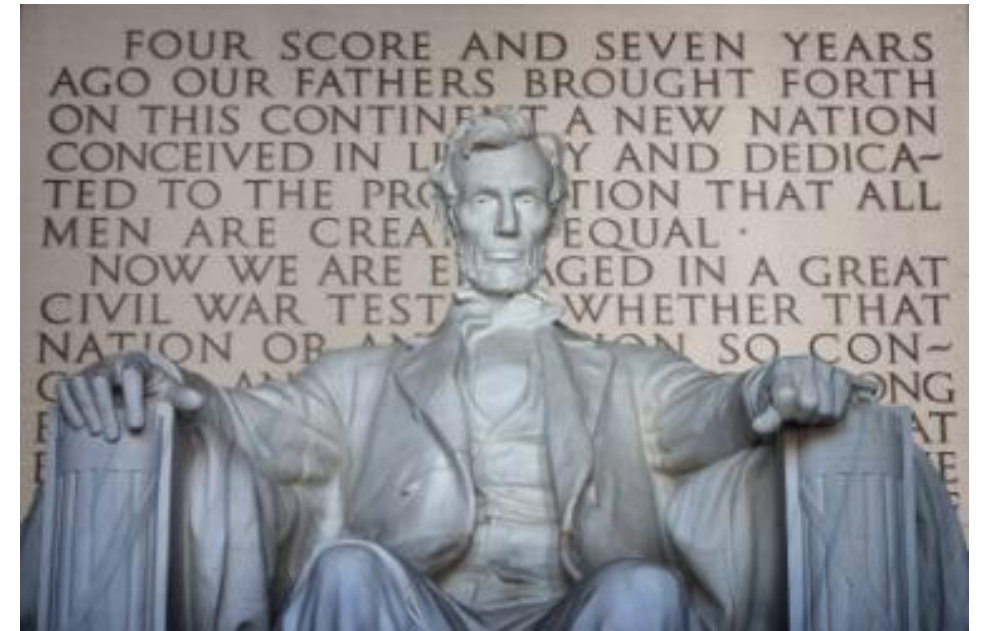
Computing confidence intervals

We will learn how to compute confidence intervals over the next few classes, but first let's review some material...

Review: sampling



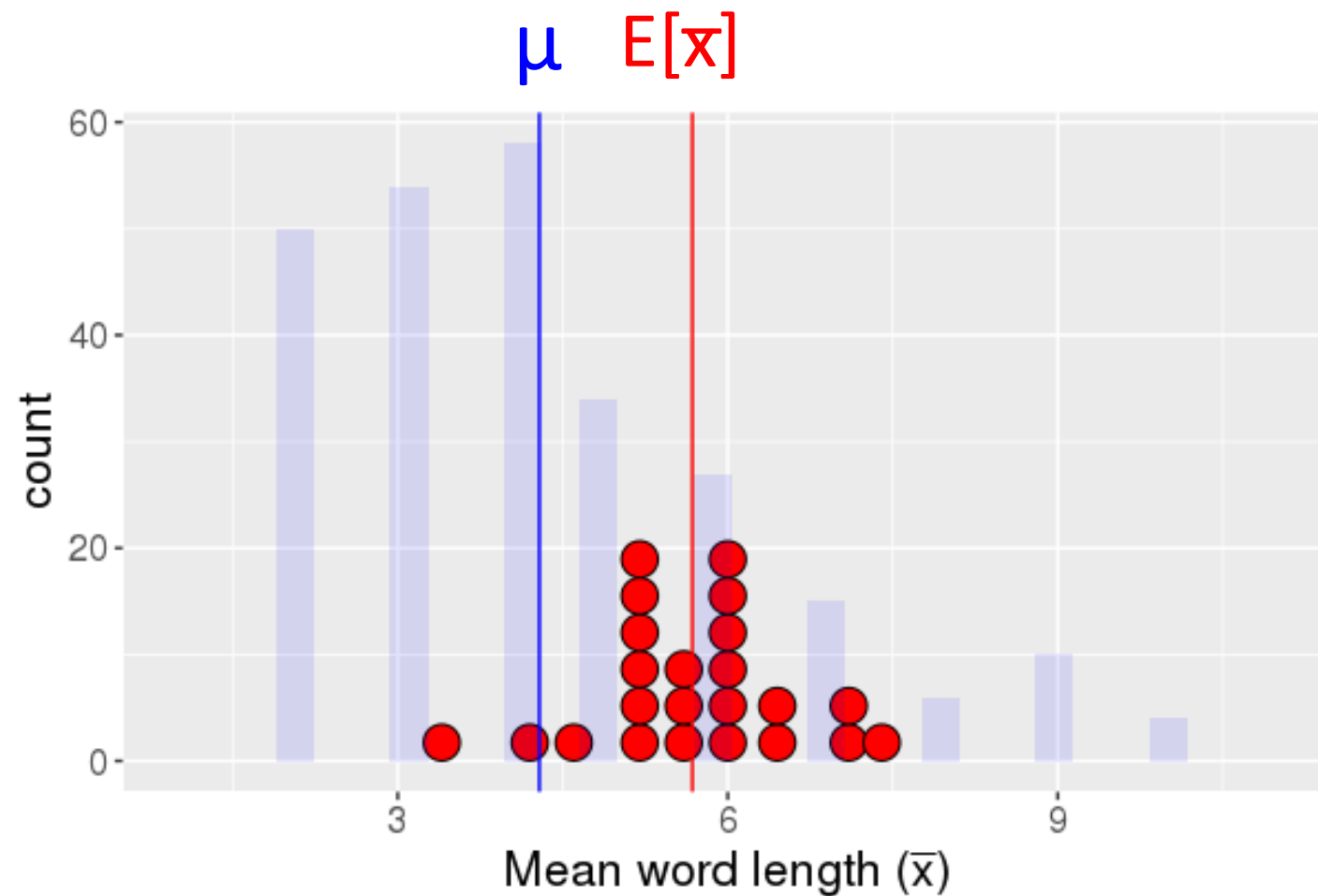
1	orange
2	red
3	green
4	white
5	white
6	white
7	white
8	white
9	red



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

A: ***n***

Bias and the Gettysburg address word length distribution

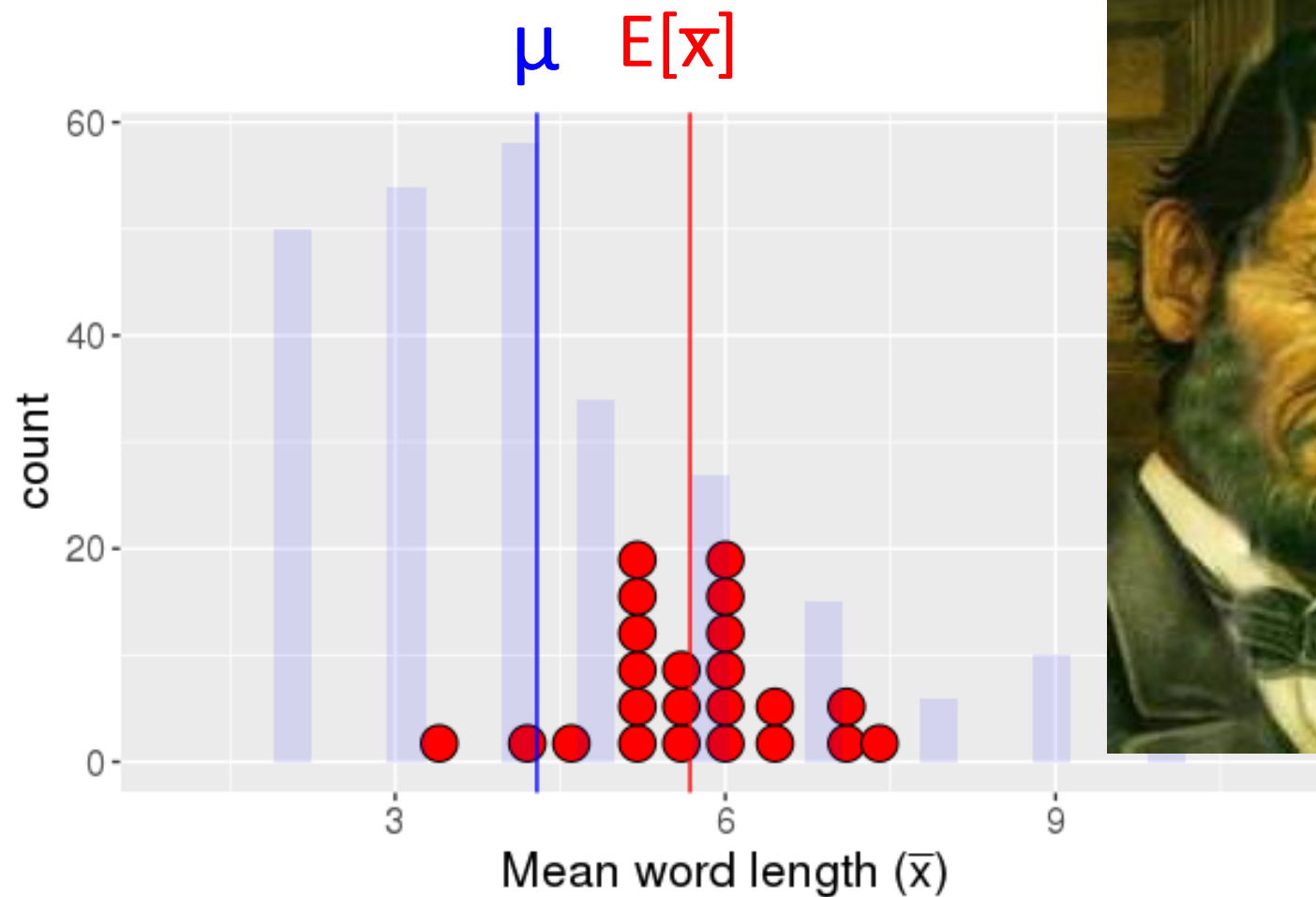


Bias and the Gettysburg address word length distribution

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

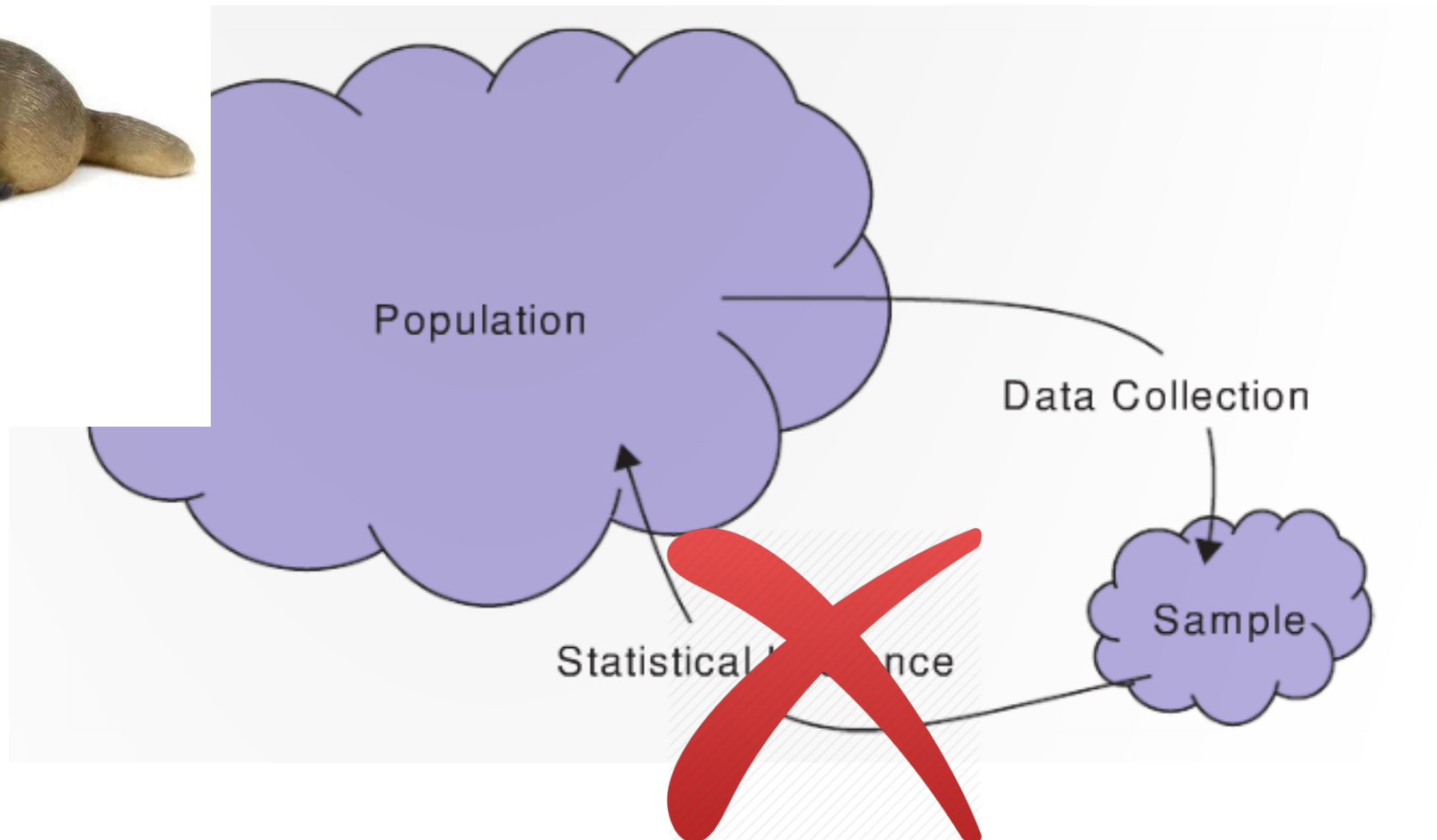
Here:

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

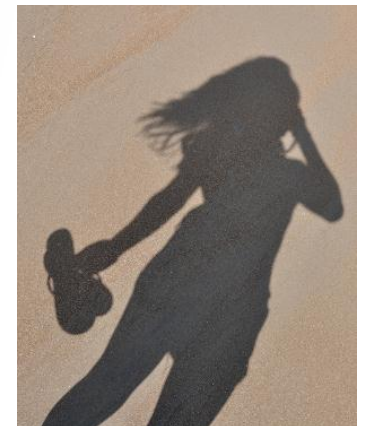


Statistical bias

μ



$\bar{x}$



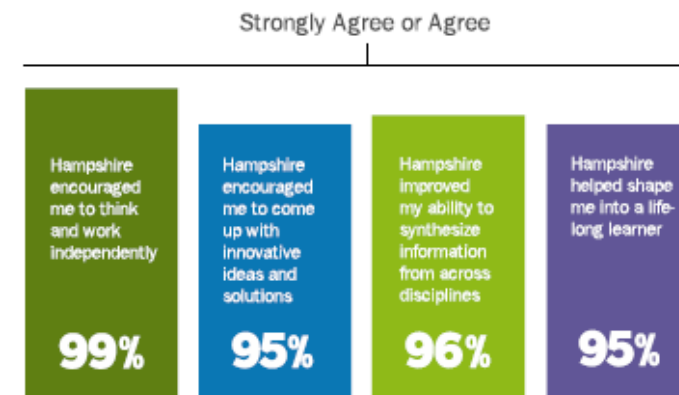
Bias or No Bias?

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%.

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?



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.

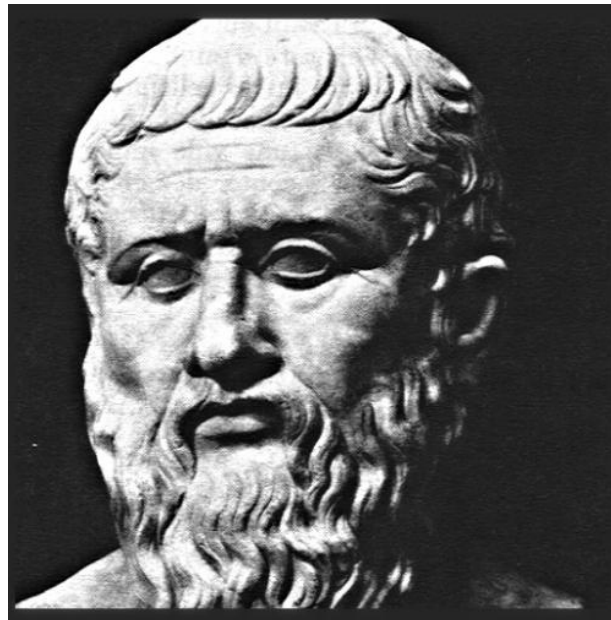
Hampshire has encouraged me to be more engaged, socially aware and more of a critical thinker than my peers.

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?

$$\pi_{\text{replied}} \neq \pi_{\text{all}}$$

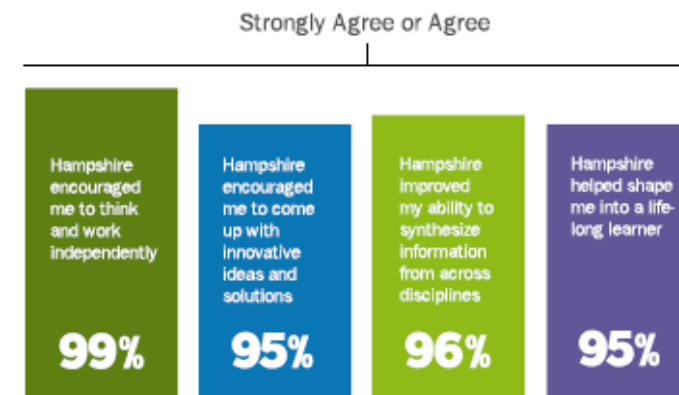


Sad Plato says:
"There's no Truth in advertising"

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?



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.

”

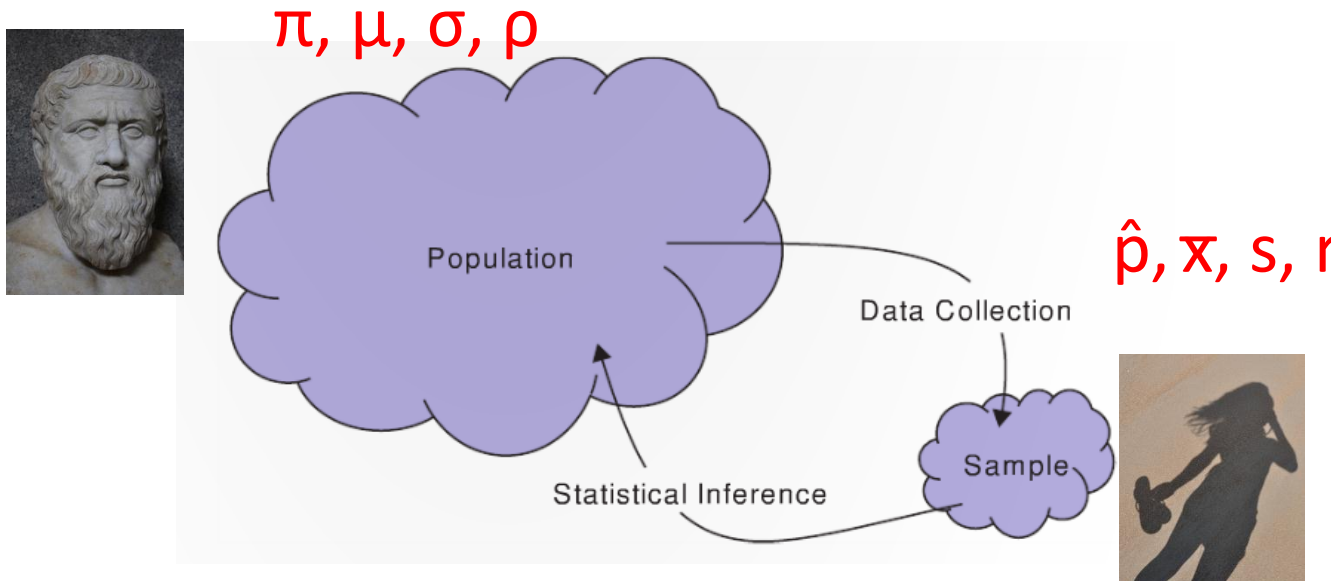
How can we prevent bias?

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

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

Allows for generalizations to the population!

Soup analogy!



Q: How do we select a random sample?

Mechanically:

- Flip coins

- Pull balls from well mixed bins

- Deal out shuffled cards, etc.

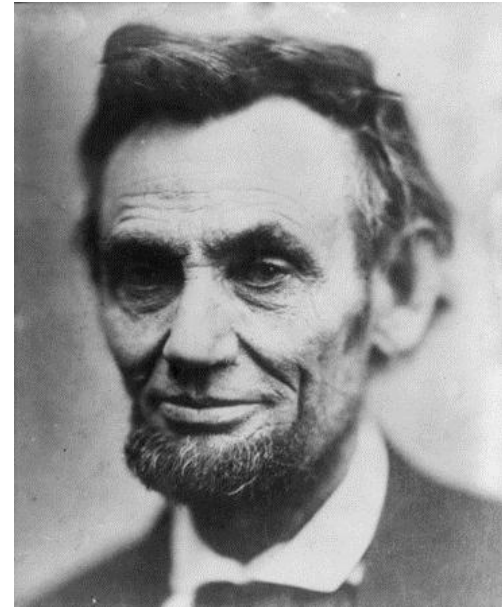
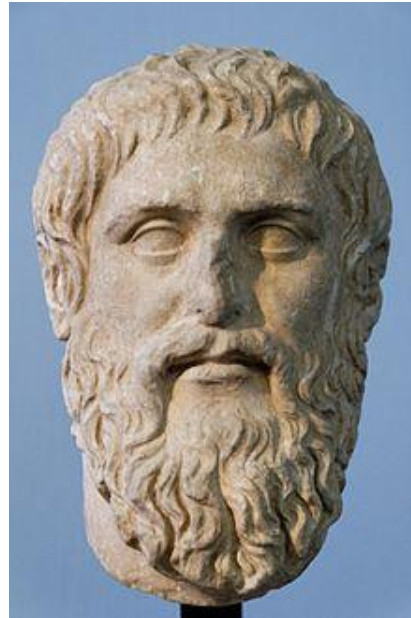
Use computer programs

Q: What computer program can we use?



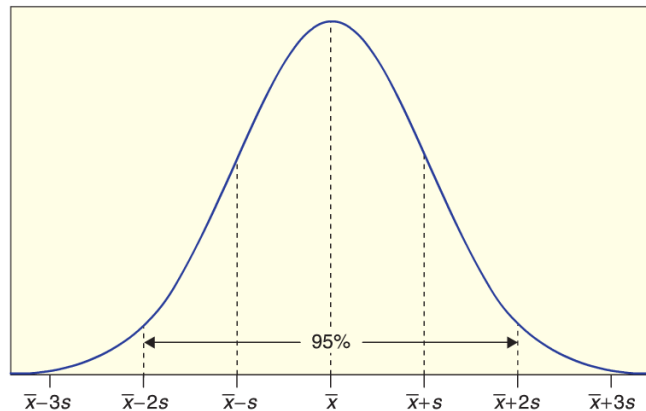
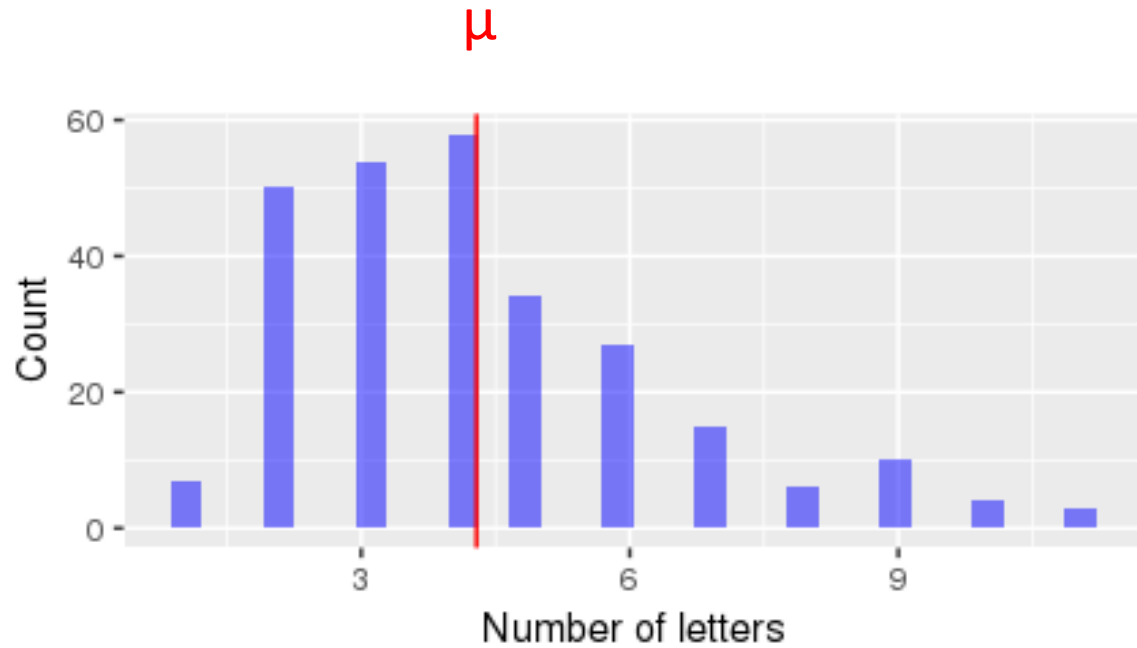
From now on we are going to assume no bias!

Happy Plato and Lincoln

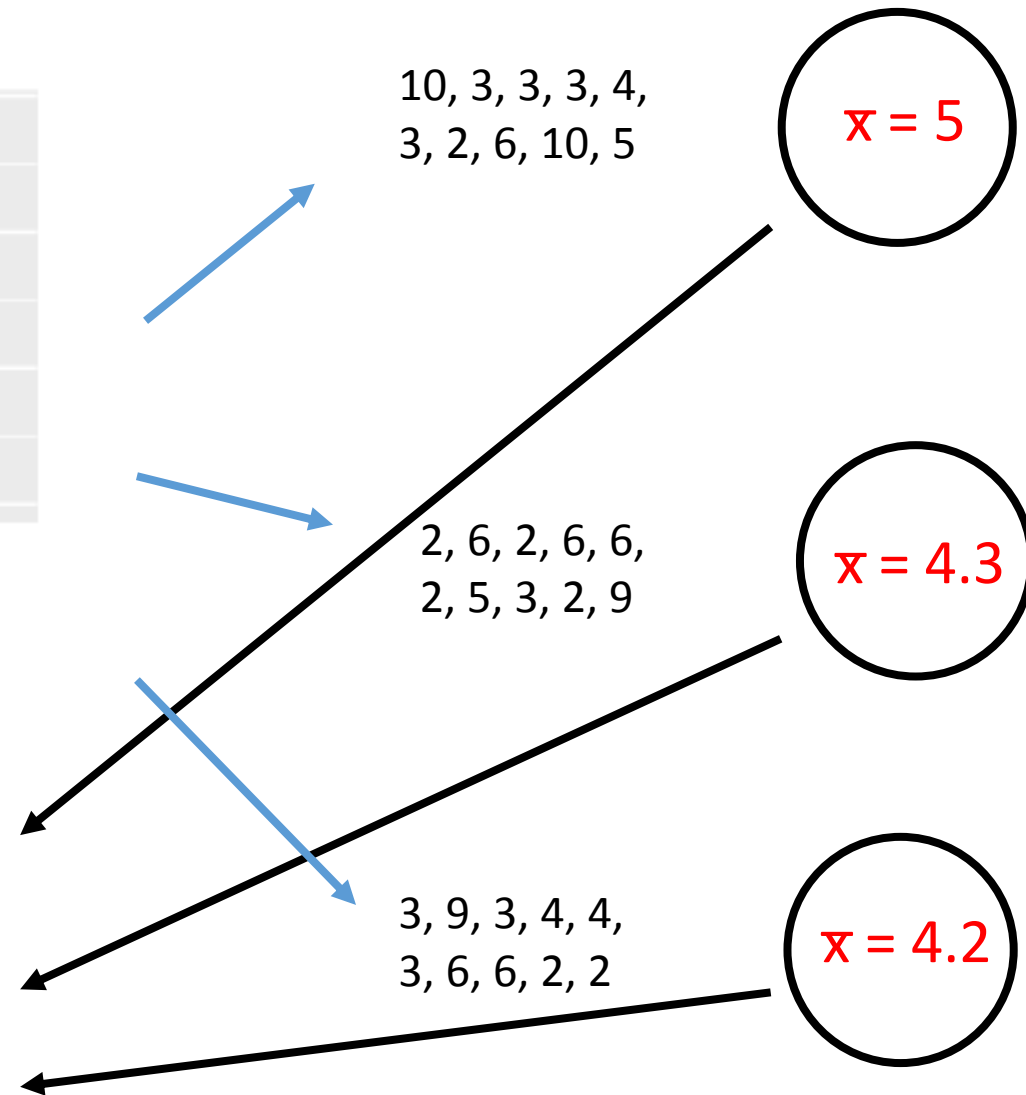


statistics, on average, reflect the parameters

Gettysburg address word length sampling distribution



Sampling distribution!



[Gettysburg sampling distribution app](#)

Sampling distribution in R

from last class and worksheet 5

```
sample_means <- NULL
```

```
for (i in 1:10000) {
```

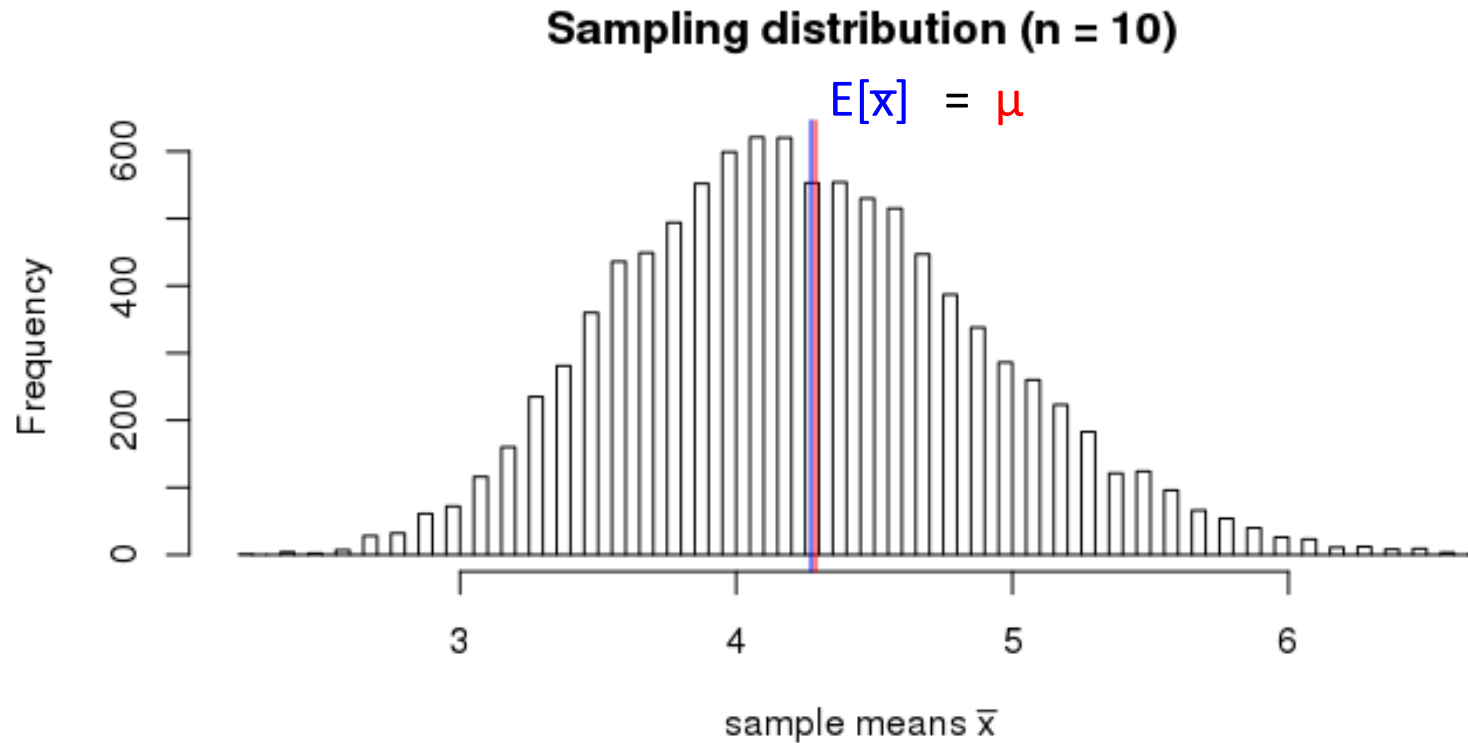
```
    curr_sample <- sample(word_lengths, 10)
```

```
    sample_means[i] <- mean(curr_sample)
```

```
}
```

```
hist(sample_means, nclass = 100)
```

Sampling distribution in R



`mean(sample_means)`

`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 $n = 1, 5, 10, 20$

```
sample_means <- NULL
```

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

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

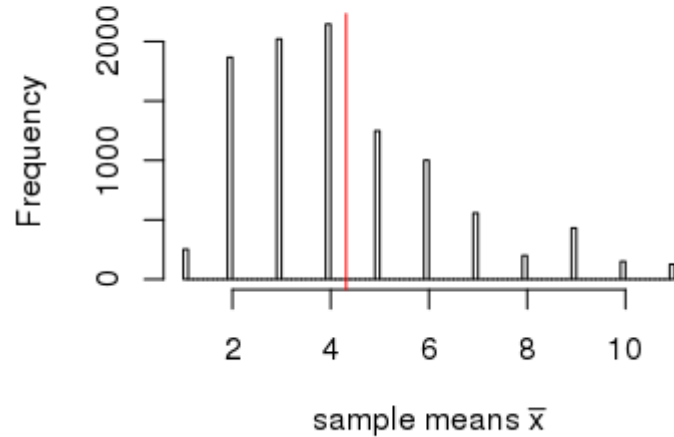
```
    sample_means[i] <- mean(curr_sample)
```

```
}
```

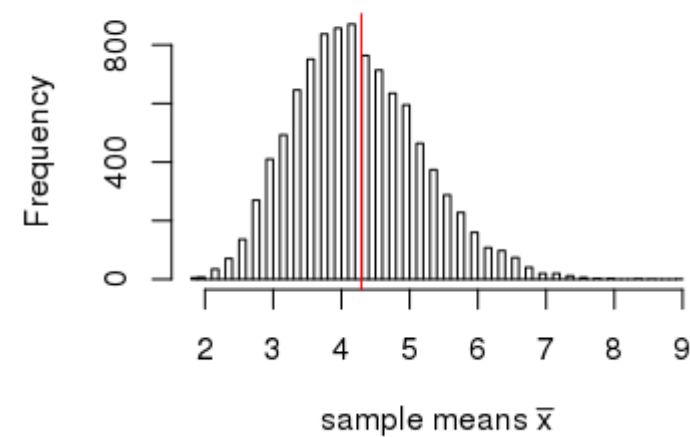
```
hist(sample_means, nclass = 100)
```

[Gettysburg sampling distribution app](#)

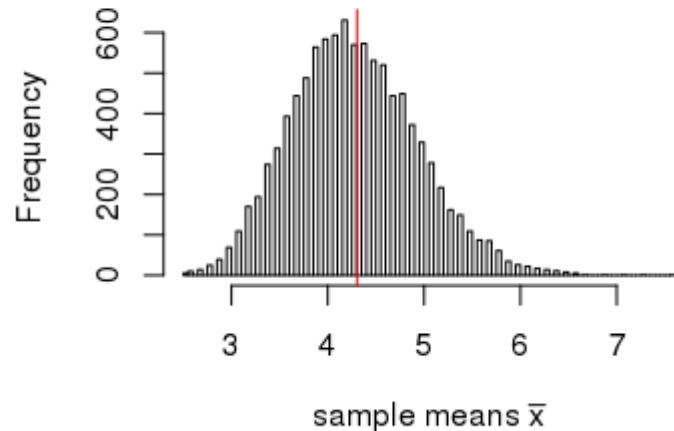
Sampling distribution (n = 1)



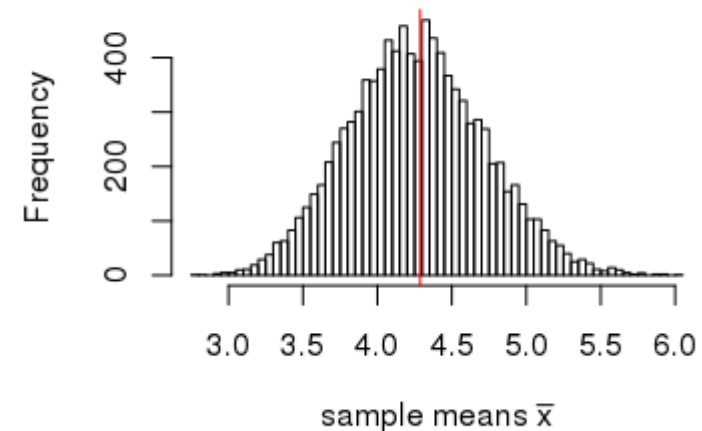
Sampling distribution (n = 5)



Sampling distribution (n = 10)



Sampling distribution (n = 20)



x-axis range 9 vs. 6

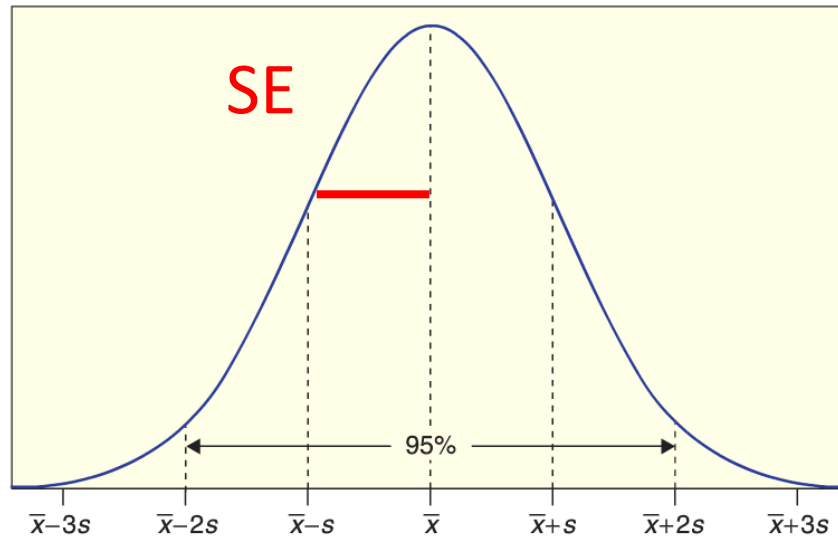
As the sample size n increases

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

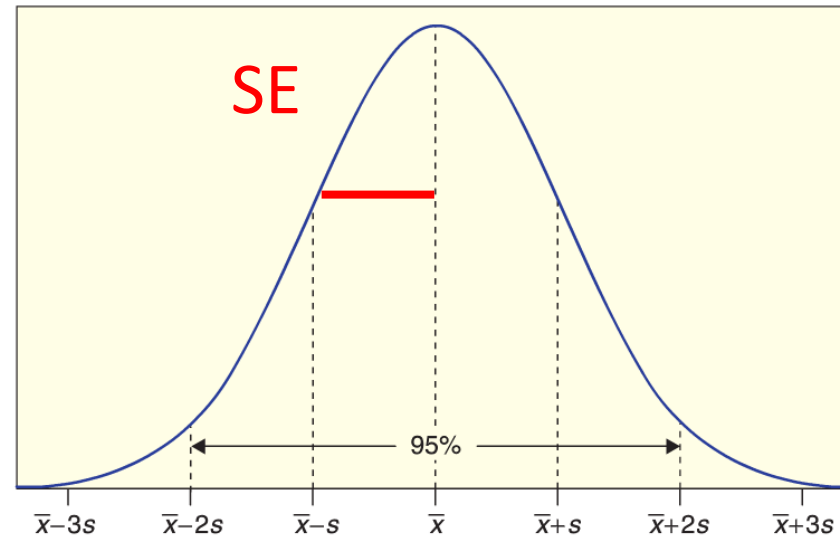
The standard error

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?



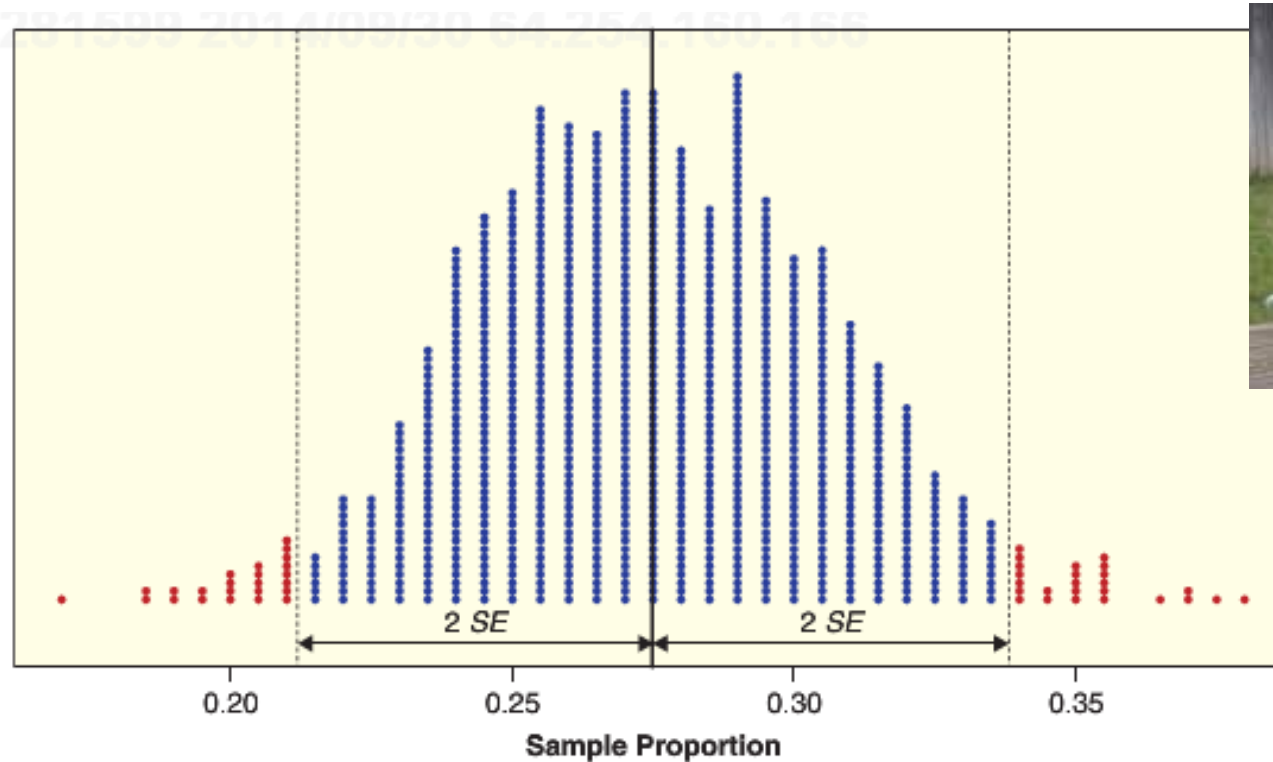
Q: If we have a large SE, would we believe a given statistic is a good estimate for the parameter?

- 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 are usually normal



So 95% of the sample statistics will fall within ± 2 Standard Errors from the mean

Thus, if we knew the SE, we could calculate a 95% confidence interval!

- Why is this true?

The problem

Unfortunately we don't know the Standard Error 😞

Q: Could we repeat the sampling process many times to estimate it?

A: No, we can't do an experiment that many times

We're just going to have to pick ourselves up from the bootstraps!

Estimate SE with $\hat{SE}$

Then use $\bar{x} \pm 2 \cdot \hat{SE}$ to get the 95% CI

Coming soon

Using the bootstrap to estimate the SE...



Read Lock 5 sections 3.1 and 3.2