Sampling distributions,
standard errors, and confidence intervals

Overview

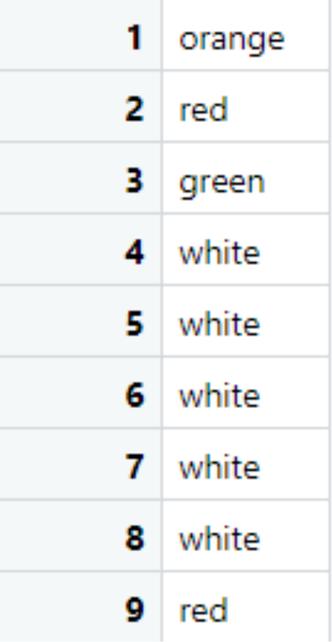
Review of sampling bias

Sampling bias and sampling distributions

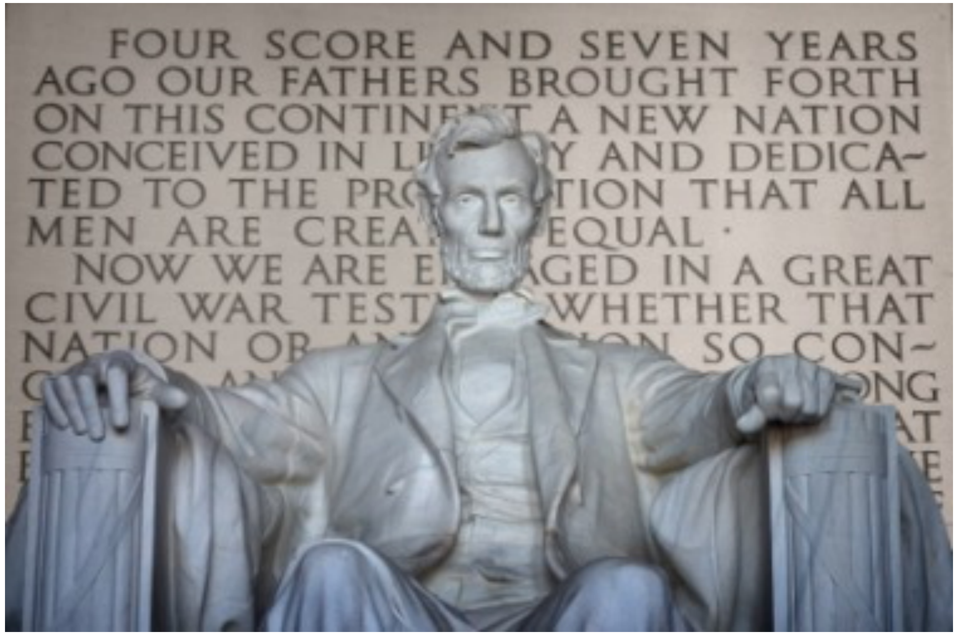
More on sampling distributions and the Standard Error

Point estimates and confidence intervals

Review: sampling



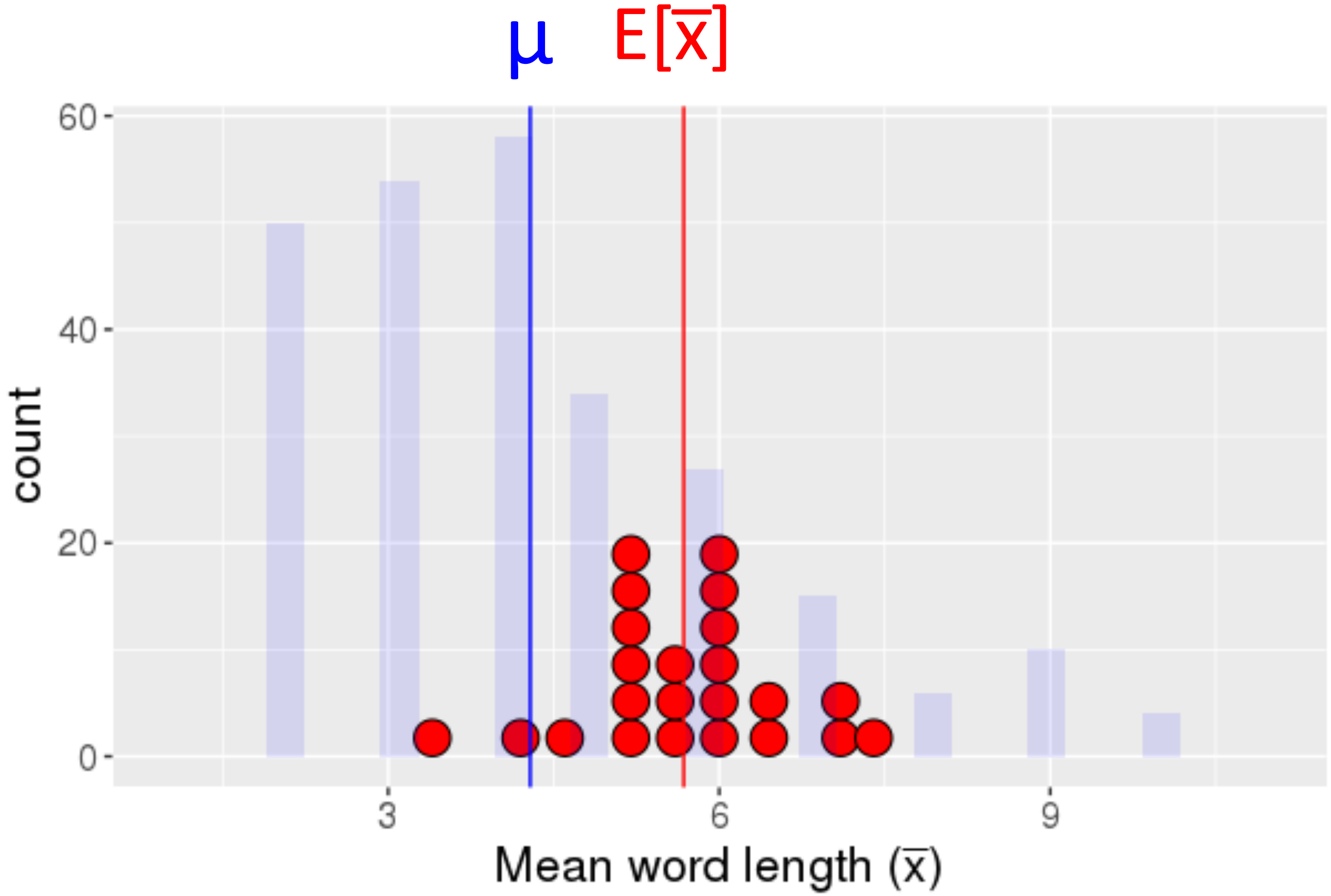
1	orange
2	red
3	green
4	white
5	white
6	white
7	white
8	white
9	red


FOUR SCORE AND SEVEN YEARS
AGO OUR FATHERS BROUGHT FORTH
ON THIS CONTIN T A NEW NATION
CONCEIVED IN LI Y AND DEDICA-
TED TO THE PR TION THAT ALL
MEN ARE CREA EQUAL ·
NOW WE ARE E GED IN A GREAT
CIVIL WAR TEST WHETHER THAT
NATION OR SO CON-

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

A: ***n***

Bias and the Gettysburg address word length distribution

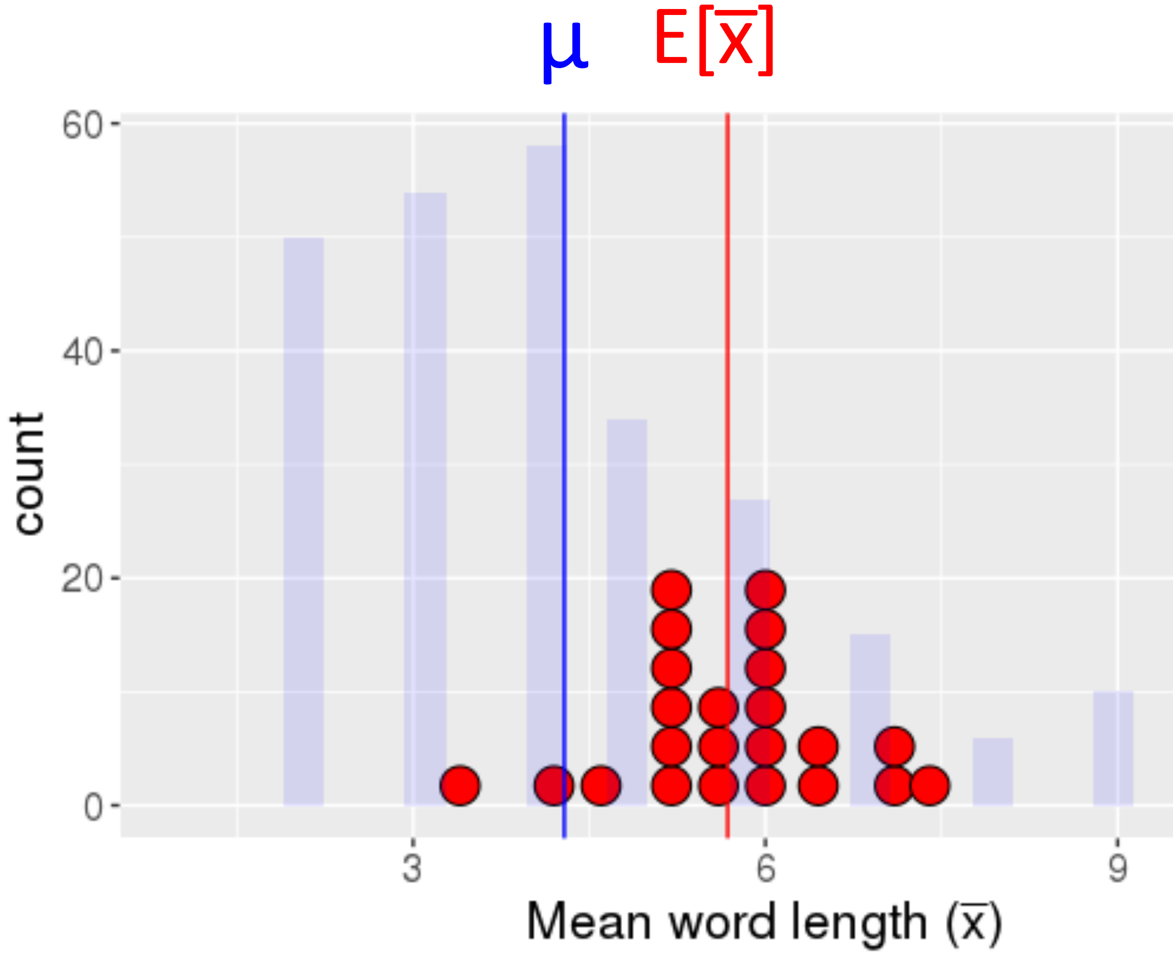


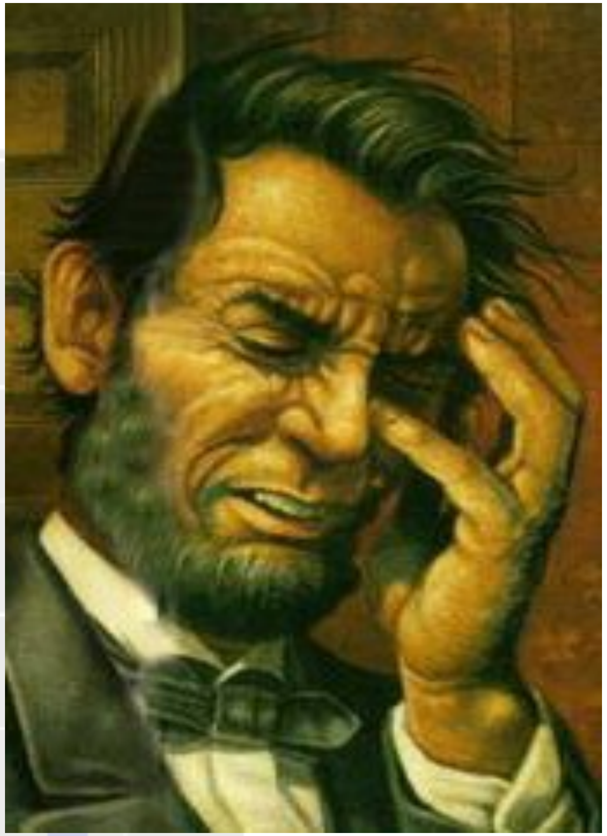
Bias and the Gettysburg address word length distribution

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

Here:

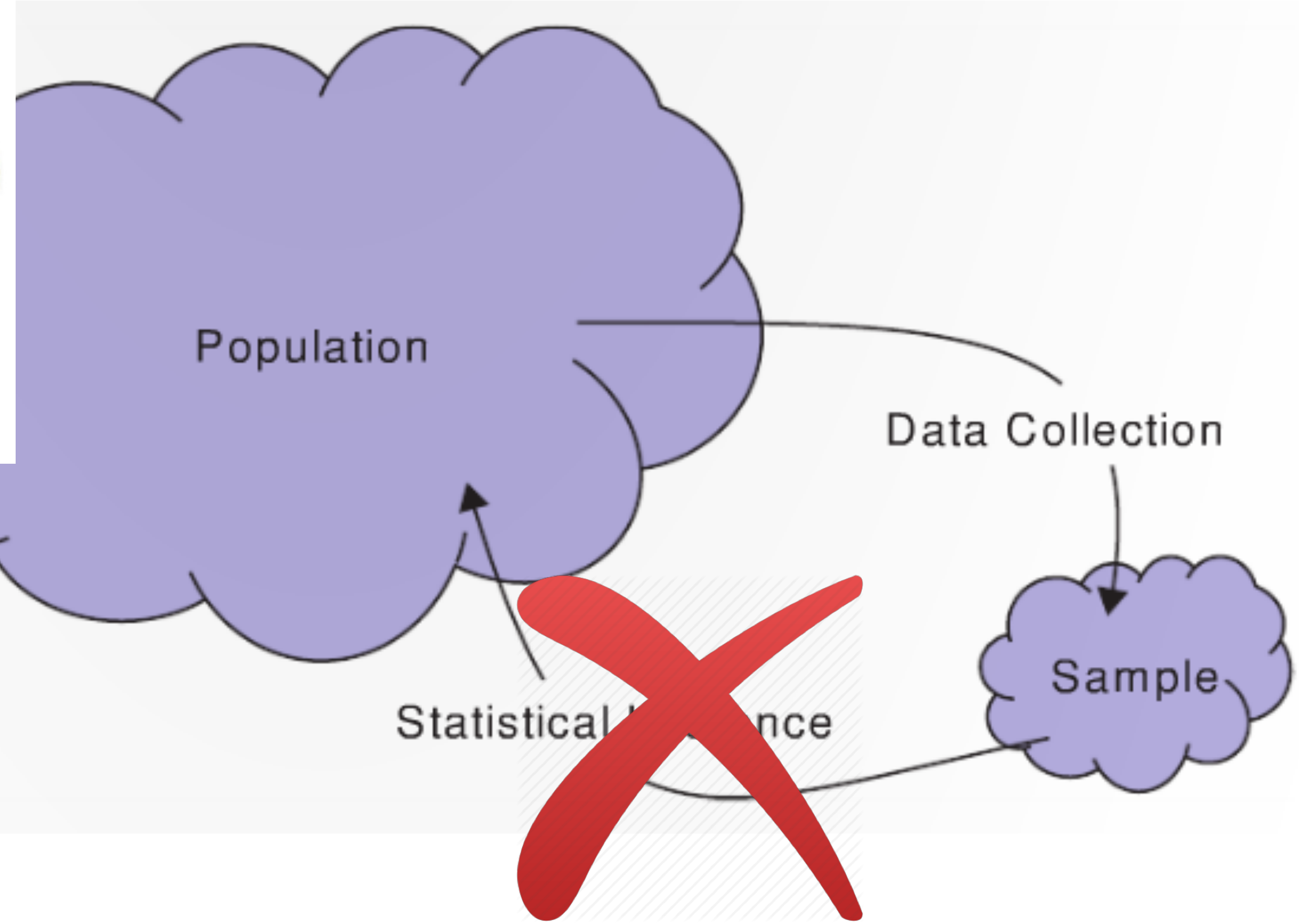
$E[\bar{x}] \neq \mu$





Statistical bias

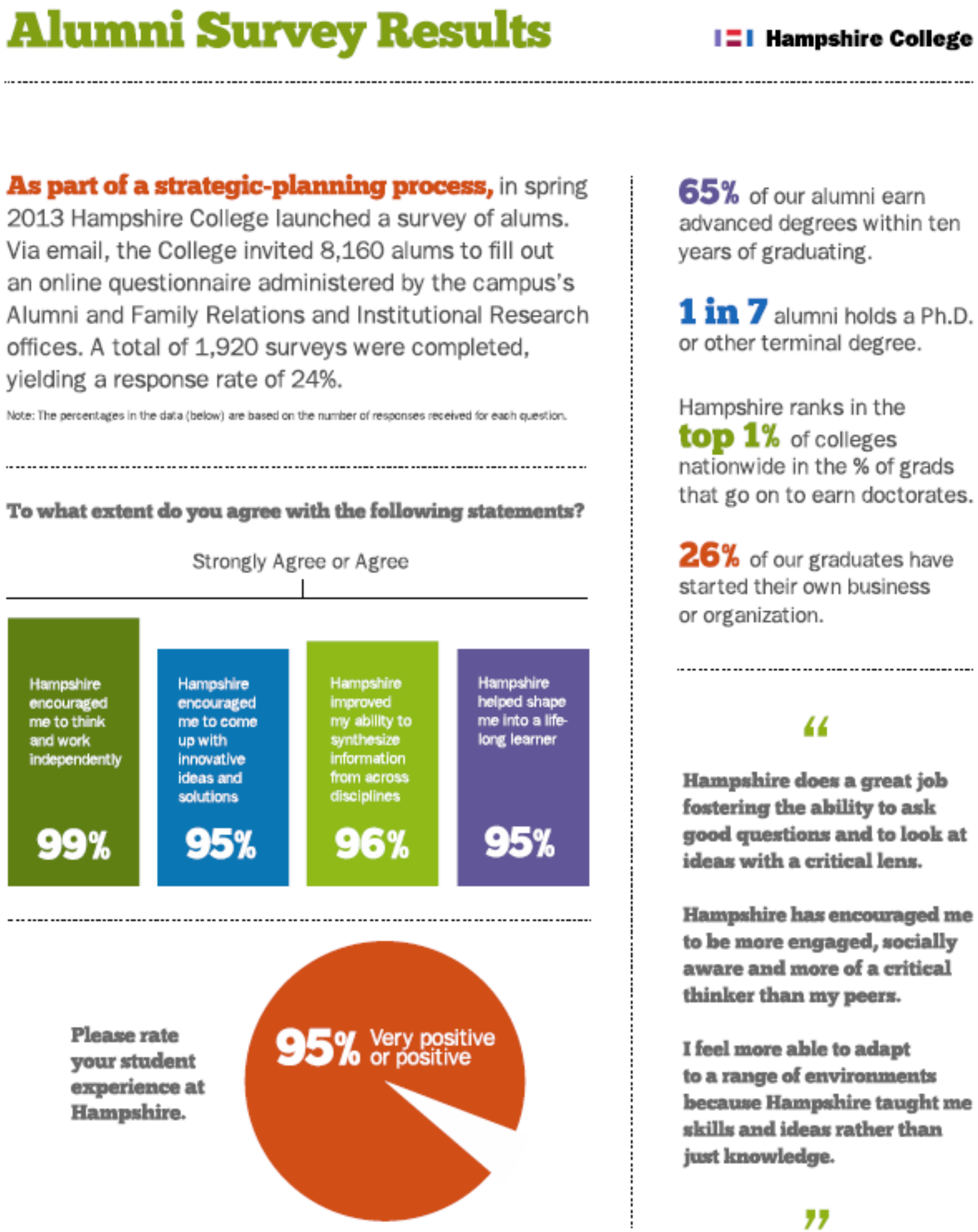
μ



$\bar{x}$

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%.**



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

Statement	Strongly Agree or Agree
Hampshire encouraged me to think and work independently	99%
Hampshire encouraged me to come up with innovative ideas and solutions	95%
Hampshire improved my ability to synthesize information from across disciplines	96%
Hampshire helped shape me into a life-long learner	95%

Please rate your student experience at Hampshire.

95% Very positive or positive

65% of our alumni earn advanced degrees within ten years of graduating.

1 in 7 alumni holds a Ph.D. or other terminal degree.

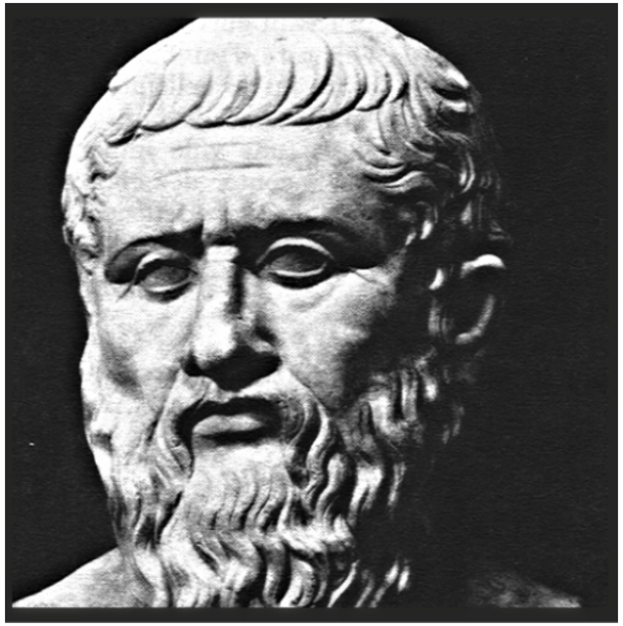
Hampshire ranks in the **top 1%** of colleges nationwide in the % of grads that go on to earn doctorates.

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.**”

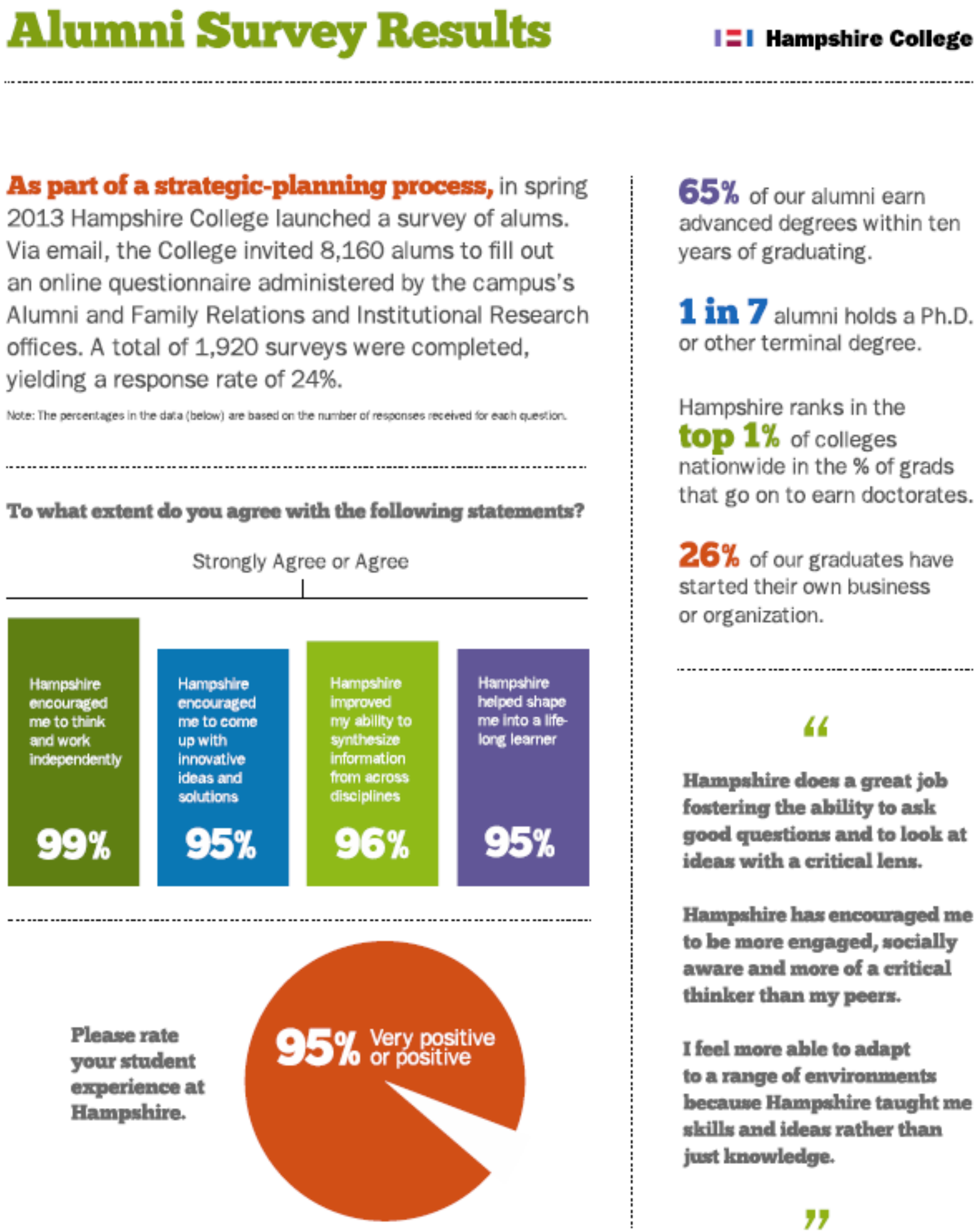
Bias or No Bias?

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



Sad Plato says:

"There's no Truth in advertising"



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

Hampshire encouraged me to think and work independently	Hampshire encouraged me to come up with innovative ideas and solutions	Hampshire improved my ability to synthesize information from across disciplines	Hampshire helped shape me into a life-long learner
99%	95%	96%	95%

Please rate your student experience at Hampshire.

95% Very positive or positive

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.**"

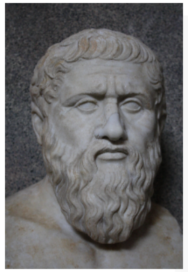
Q: How can we prevent bias?

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

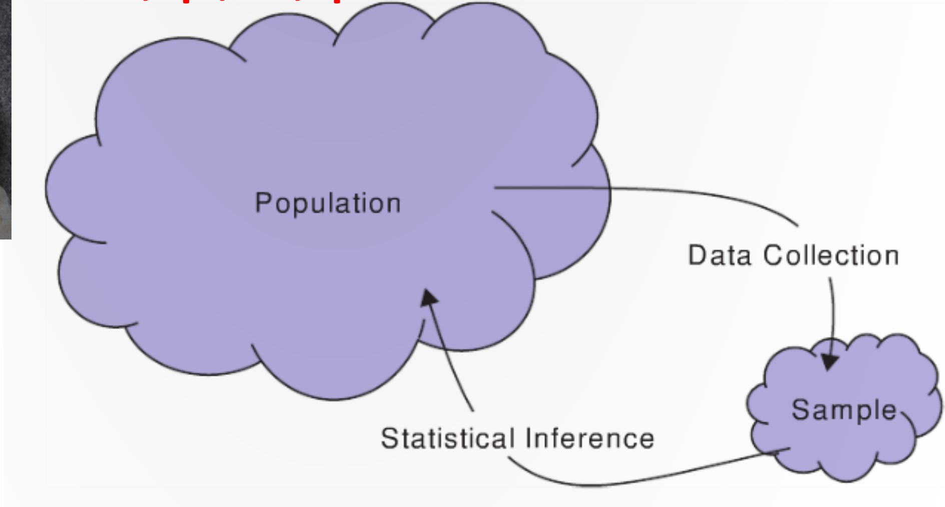
- where each member in the population is equally likely to be in the sample

This allows for generalizations to the population!

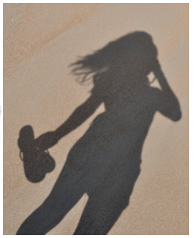
Soup analogy!



π, μ, σ, ρ



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

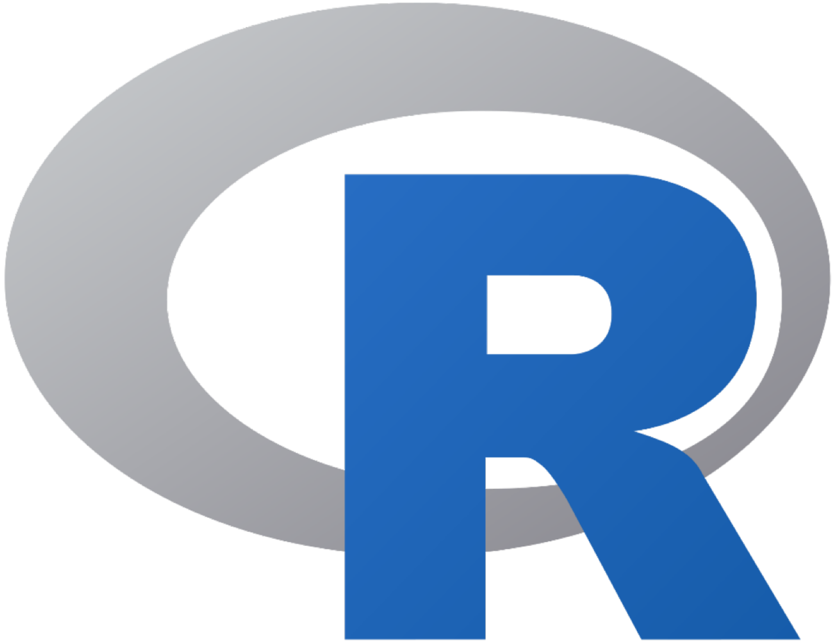


Q: How do we select a random sample?

Mechanically:

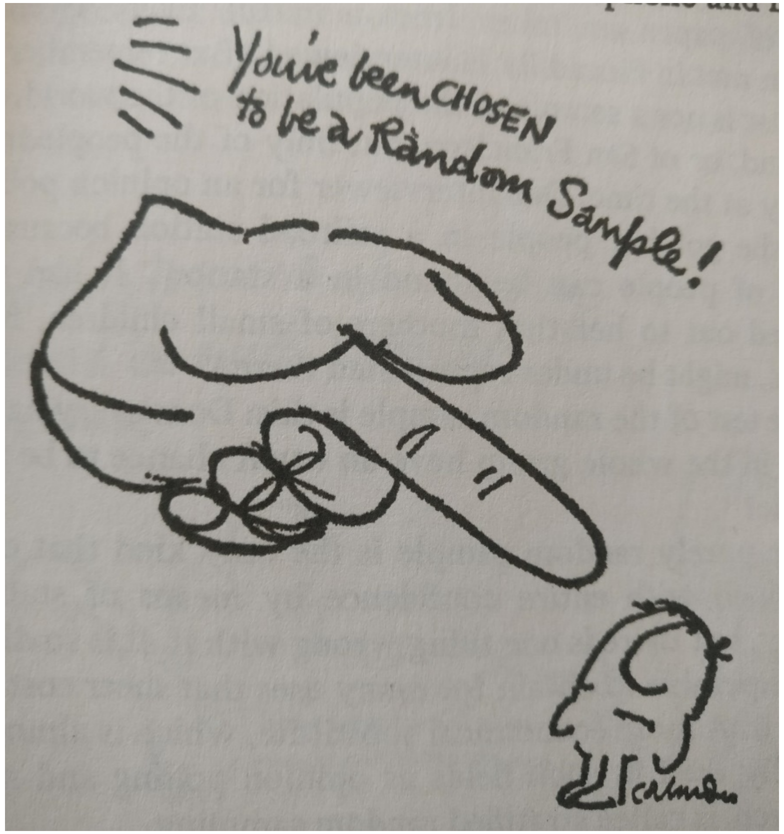
- Flip coins
- Pull balls from well mixed bins
- Deal out shuffled cards, etc.

Use computer programs



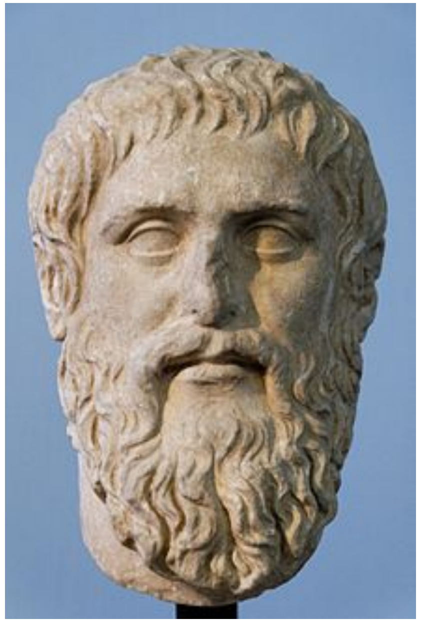
Q: What computer program can we use?

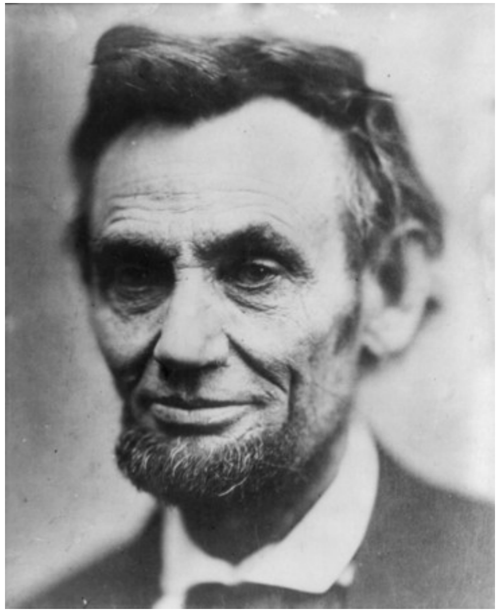
Questions about statistical bias?



From now on we are going to assume no bias!

Happy Plato and Lincoln

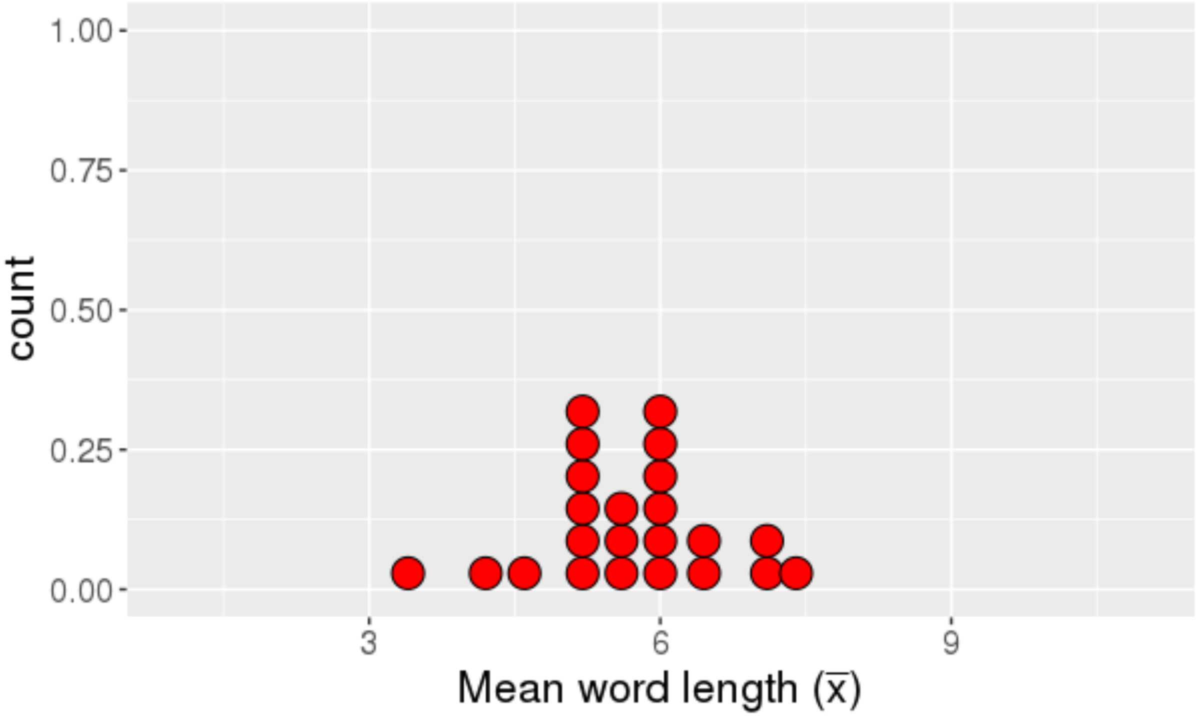




statistics, on average, reflect the parameters

For our distribution of Gettysburg word lengths…

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



A: The mean length of 10 words (x̄)

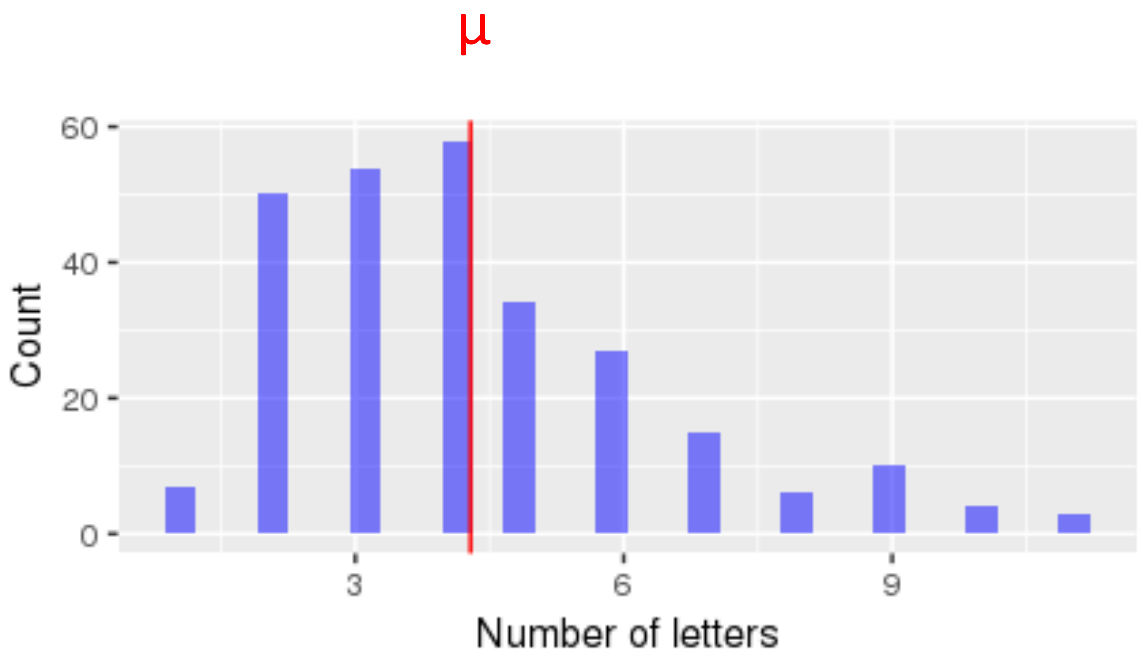
i.e., each point in our **distribution** is a statistic!

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

Gettysburg address word length sampling distribution



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

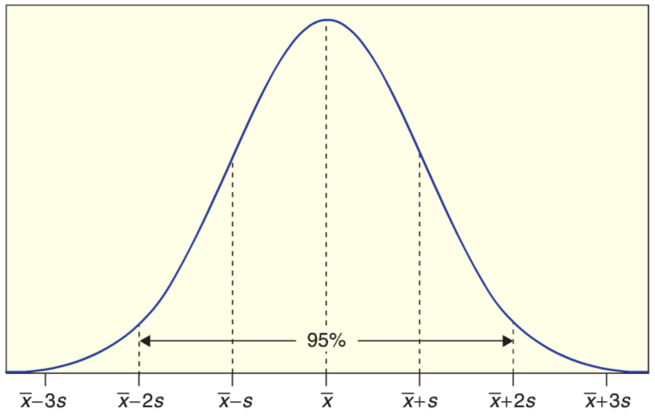
$\bar{x} = 5$

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

$\bar{x} = 4.3$

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

$\bar{x} = 4.2$



Sampling distribution!

Gettysburg sampling distribution app

Let's create a sampling distribution in R

Log into Class workspace 2 – link is on Canvas

- Link is on Canvas
- > library(ClassTools)

Get the Gettysburg population data

```
> download_class_data("gettysburg.Rda")
> load("gettysburg.Rda")
> word_lengths <- gettysburg$num_letters
```

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

```
do_it(100)  *  {

        2 + 3

}
```

Let's create a sampling distribution in R

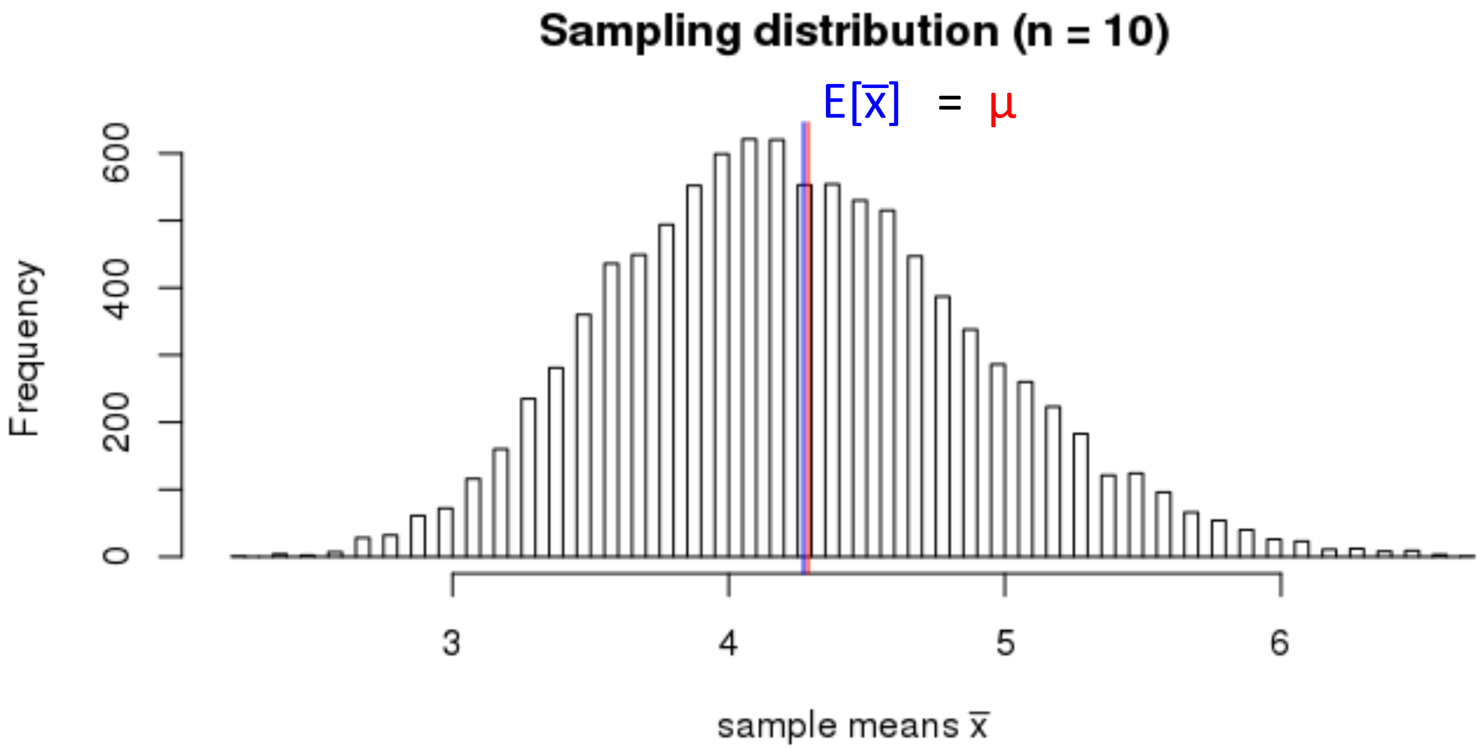
```
sampling_dist <- do_it(10000)  *  {

        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

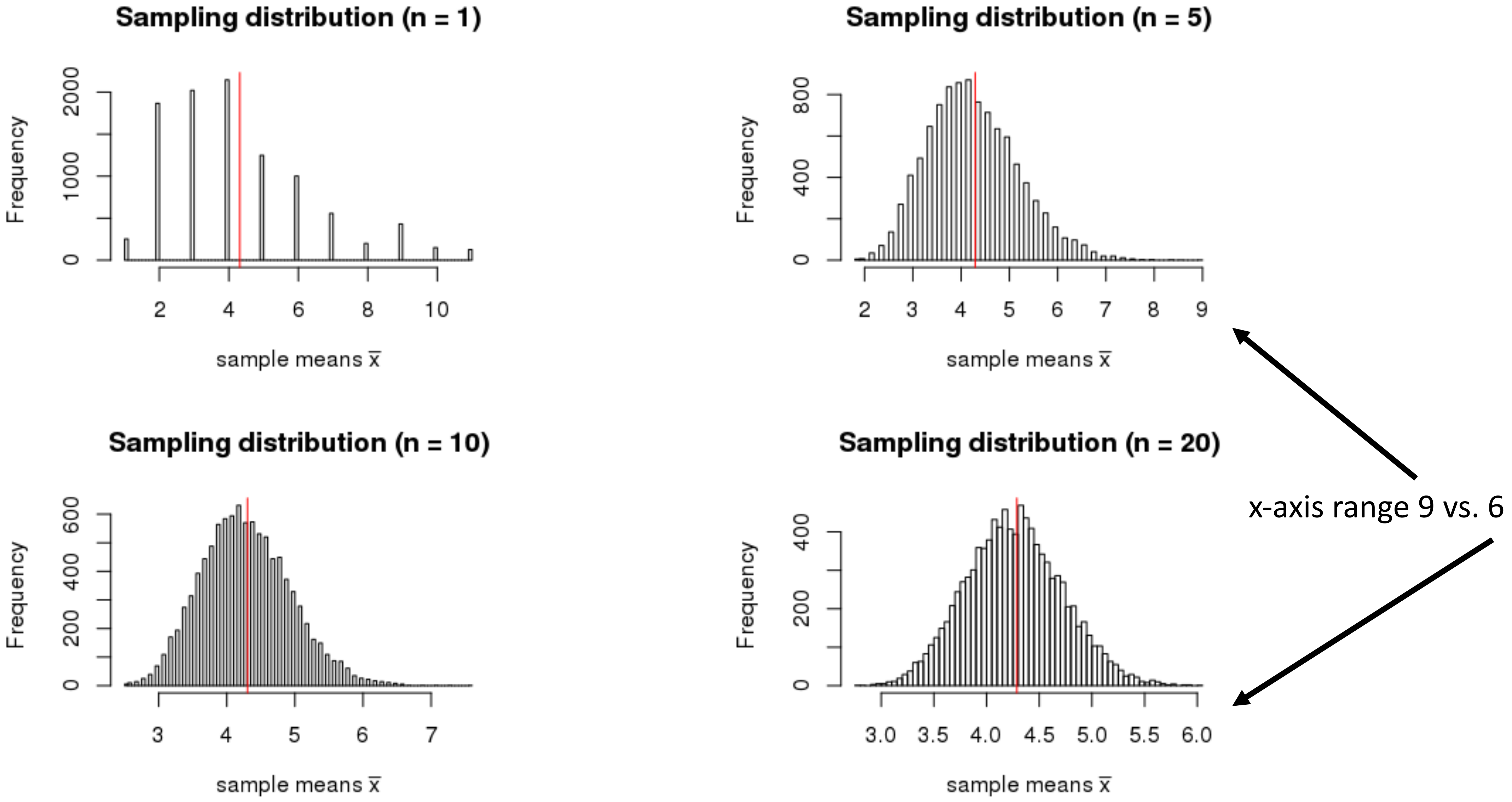
```
sampling_dist <- do_it(10000)  *  {

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

}

hist(sample_means, 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

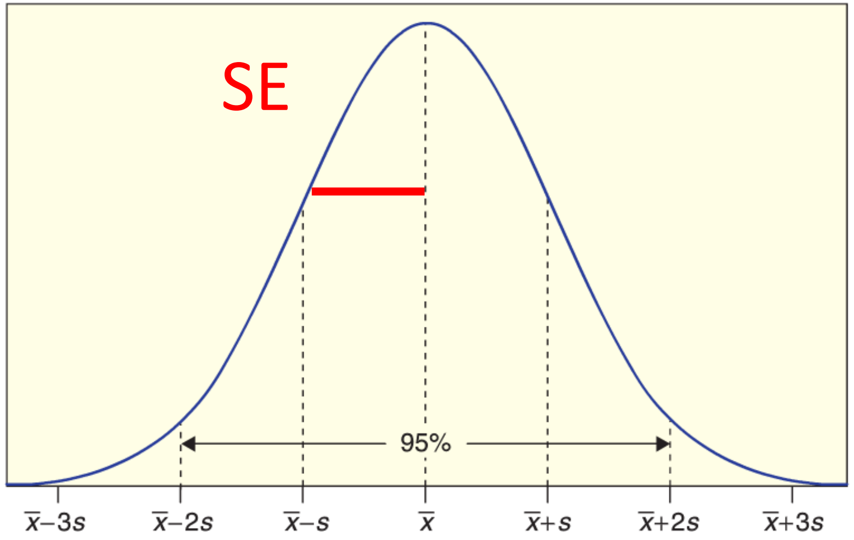
As the sample size n increases

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

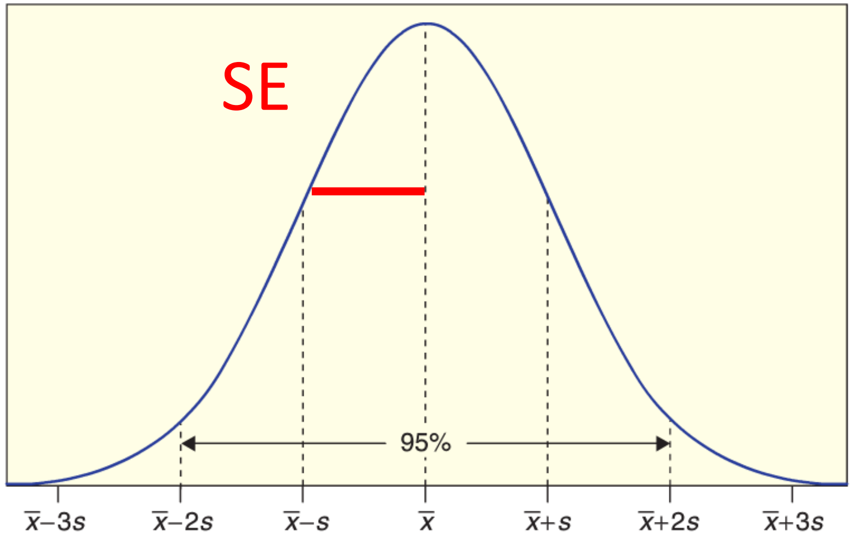
The standard error

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 $\overline{x}$ is a good estimate for μ?

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

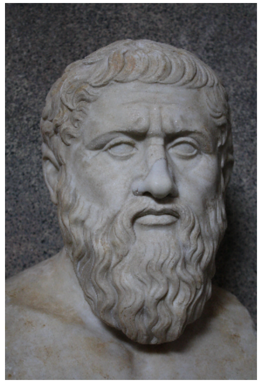
- E.g., $\overline{x}$ could be far from μ

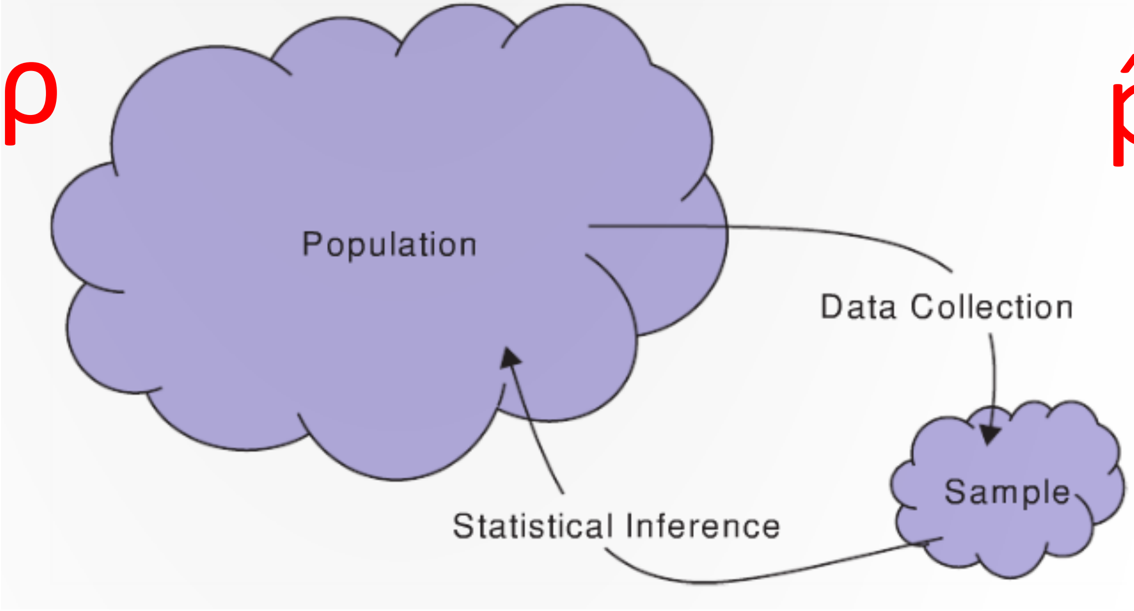
Back to the big picture: Inference

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

- $\overline{x}$ is a point estimate for…? μ

49% of American approve of Trump’s job performance according to a recent Gallup poll

Q: What are π and p̂ here?

Q: Is p̂ 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 ± 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

49% 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 46% to 52%

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

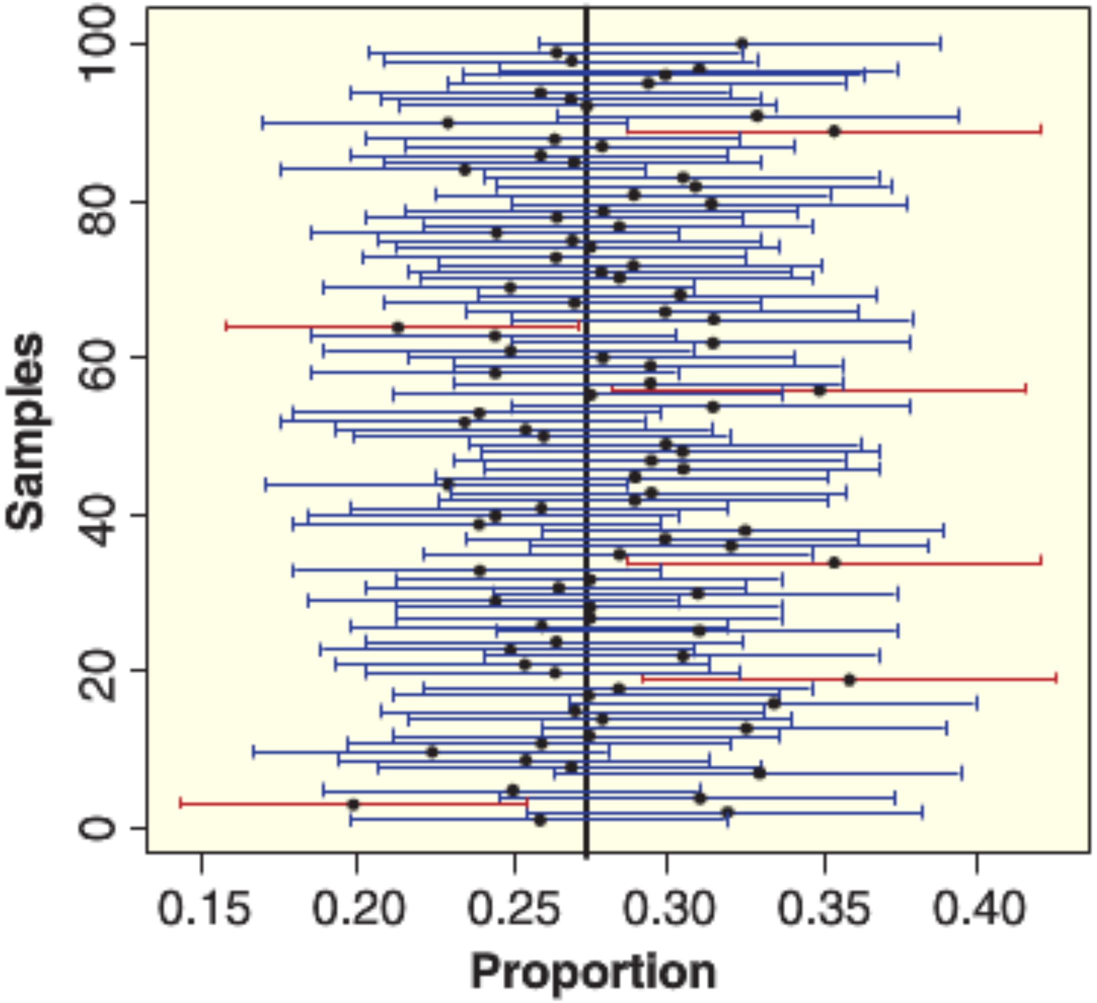
We toss intervals at it

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


Samples
Proportion