# Sampling distributions, standard errors, and confidence intervals

# Overview

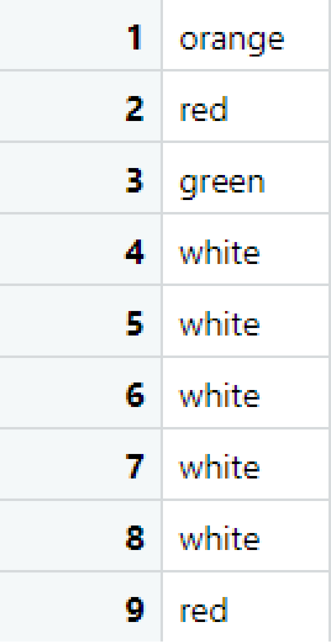
Review of bias and sampling distributions

Exploring sampling distributions in R and the Standard Error

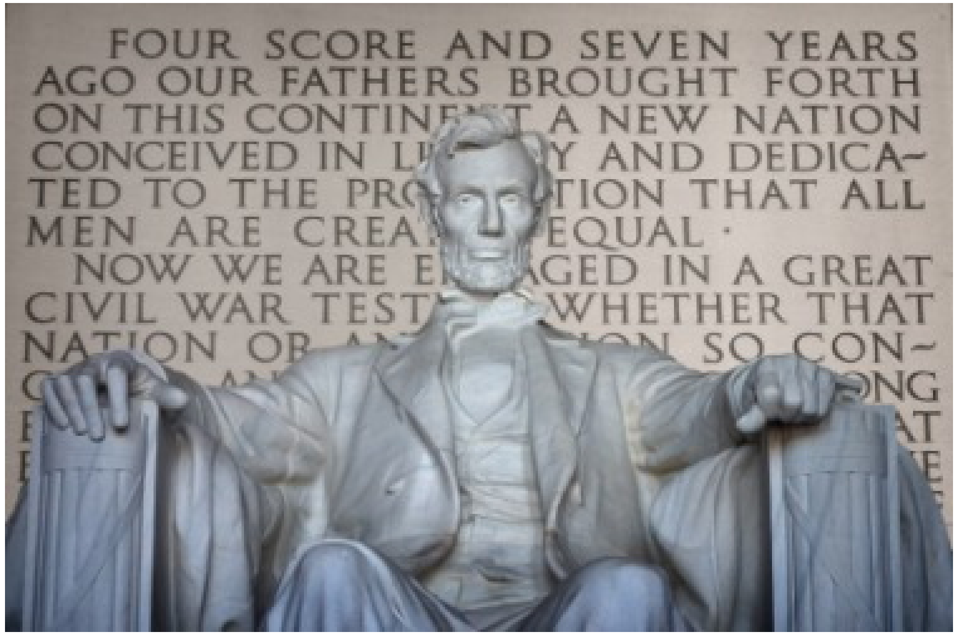
Point estimates and confidence intervals

# Review: sampling and sampling distributions

# Review: sampling



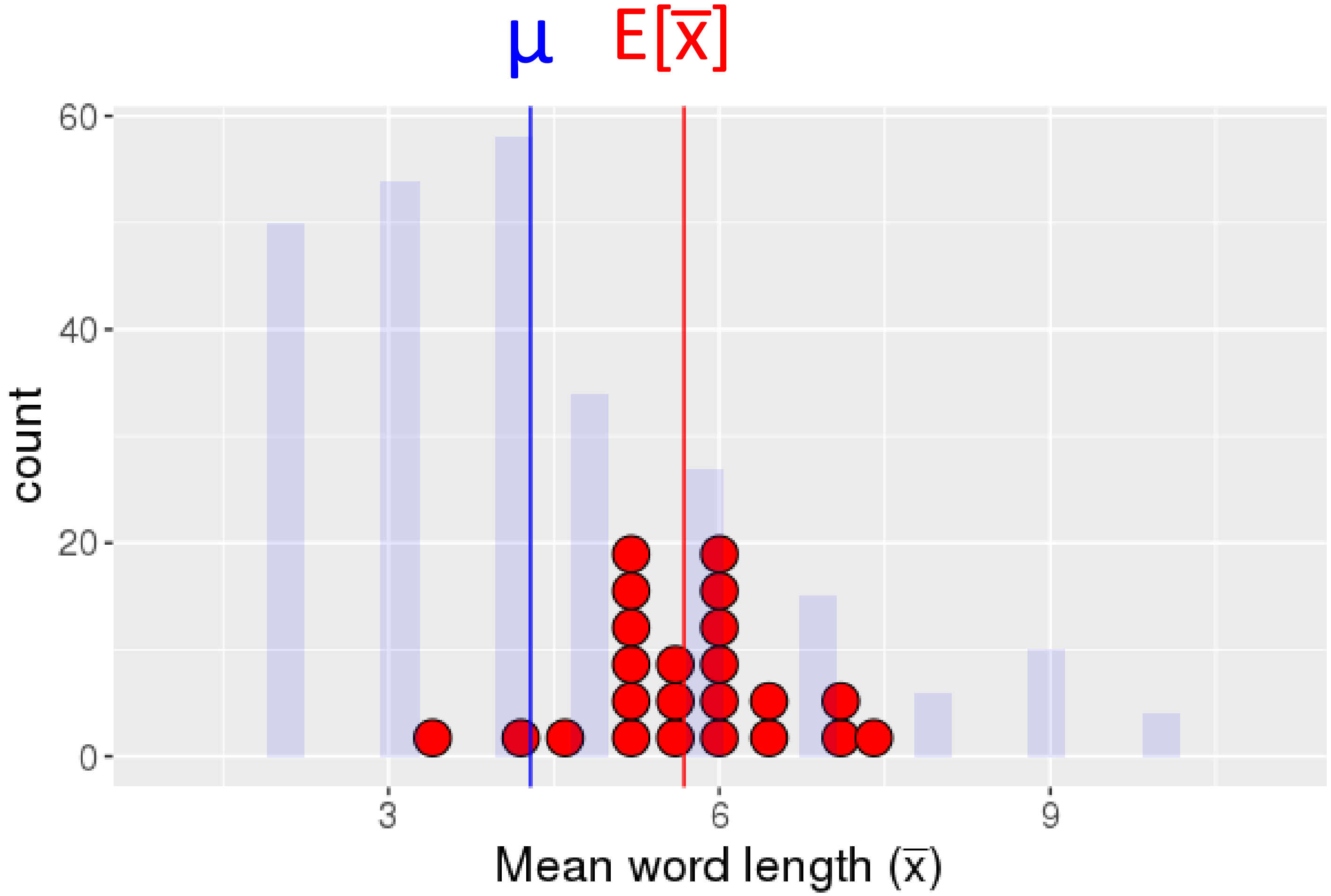
| | |
|---|---|
| 1 | orange |
| 2 | red |
| 3 | green |
| 4 | white |
| 5 | white |
| 6 | white |
| 7 | white |
| 8 | white |
| 9 | red |


FOUR SCORE AND SEVEN YEARS
AGO OUR FATHERS BROUGHT FORTH
ON THIS CONTINE 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

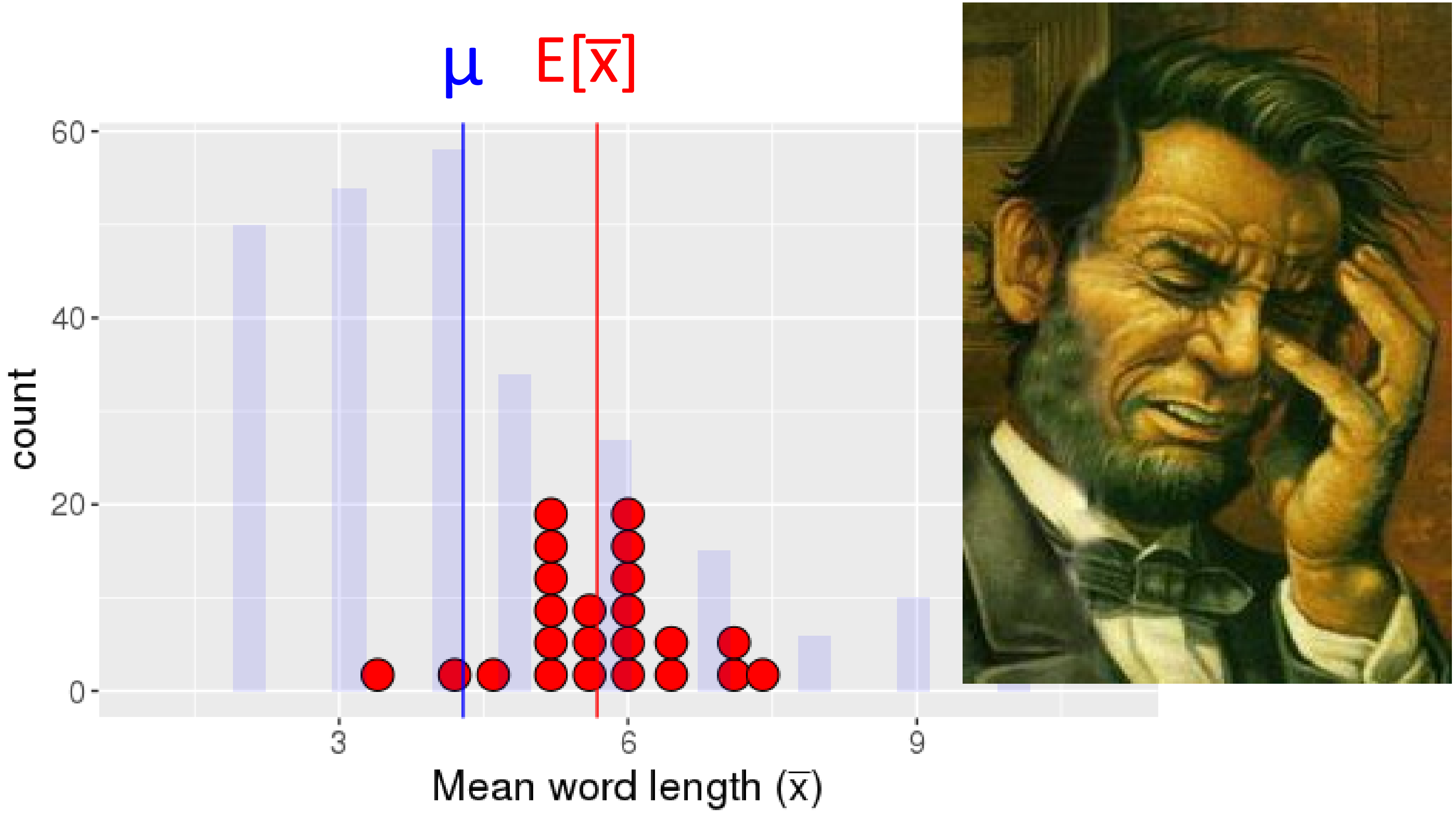


# Bias and the Gettysburg address word length distribution

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

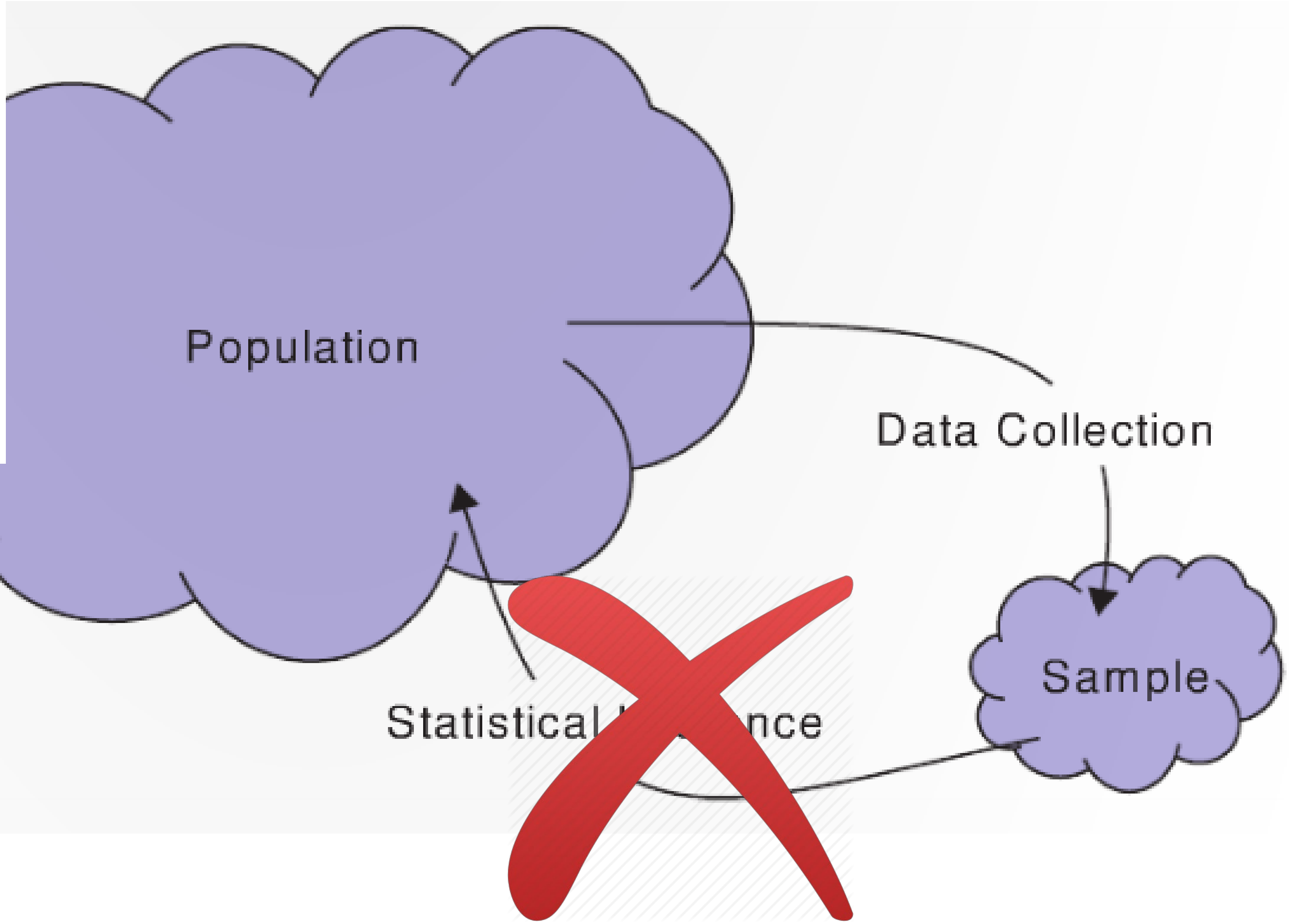
Here:

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



# Statistical bias

$\mu$

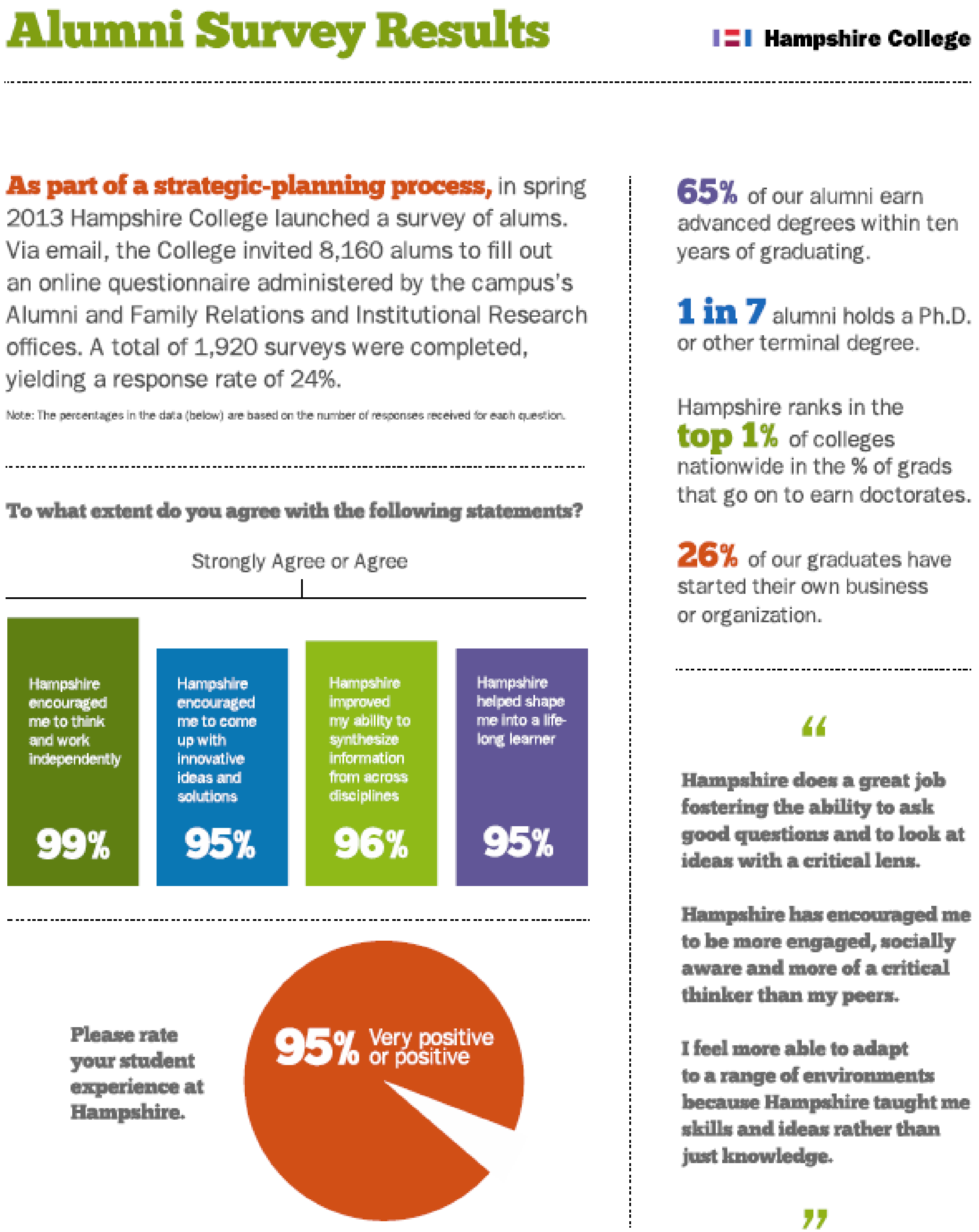


$\bar{x}$

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.

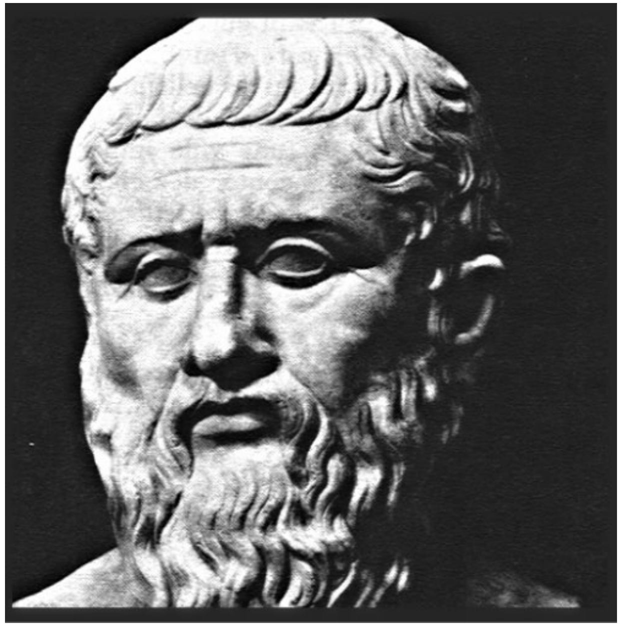
Hampshire ranks in the **top 1%** of colleges nationwide in the % of grads that go on to earn doctorates.

**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.**”

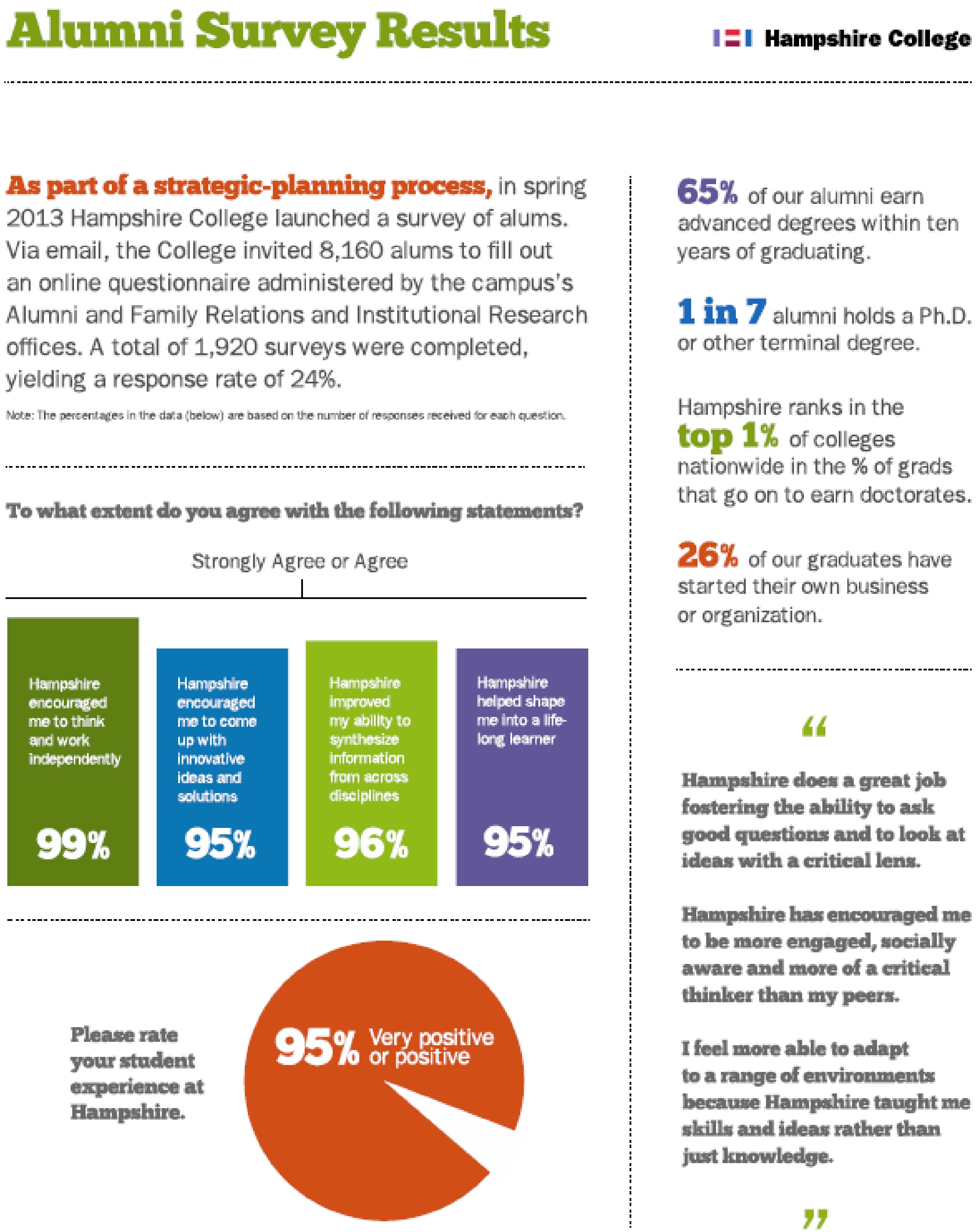
# Bias or No Bias?

$$E[\hat{p}_{\text{replied}}] \neq \pi_{\text{all}}$$



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."**

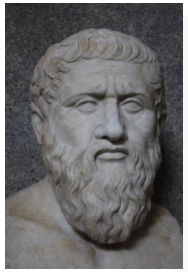
# Q: How can we prevent sampling bias?

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

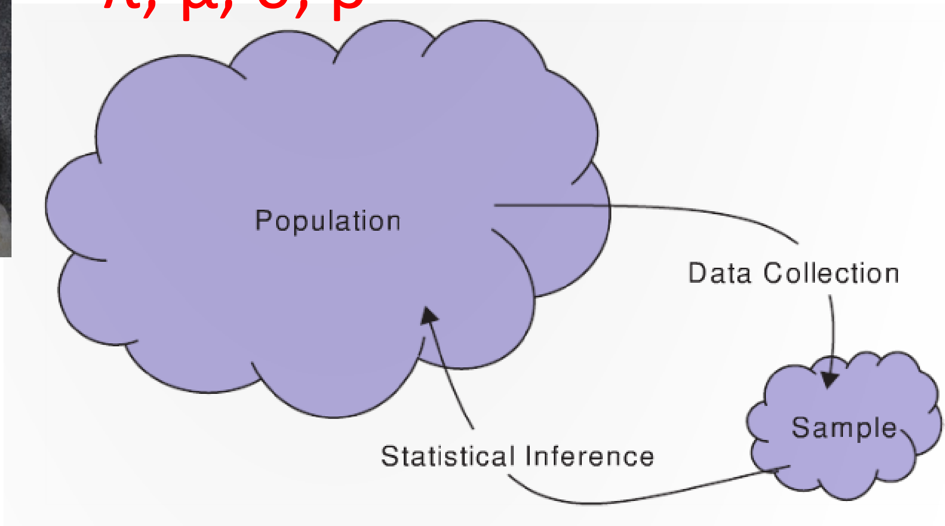
- where each member in the population is equally likely to be in the sample

This allows for generalizations to the population!

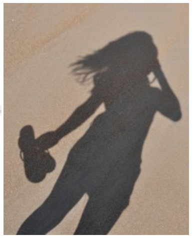
Soup analogy!



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



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



# Q: How do we select a random sample?

Mechanically:

Flip coins

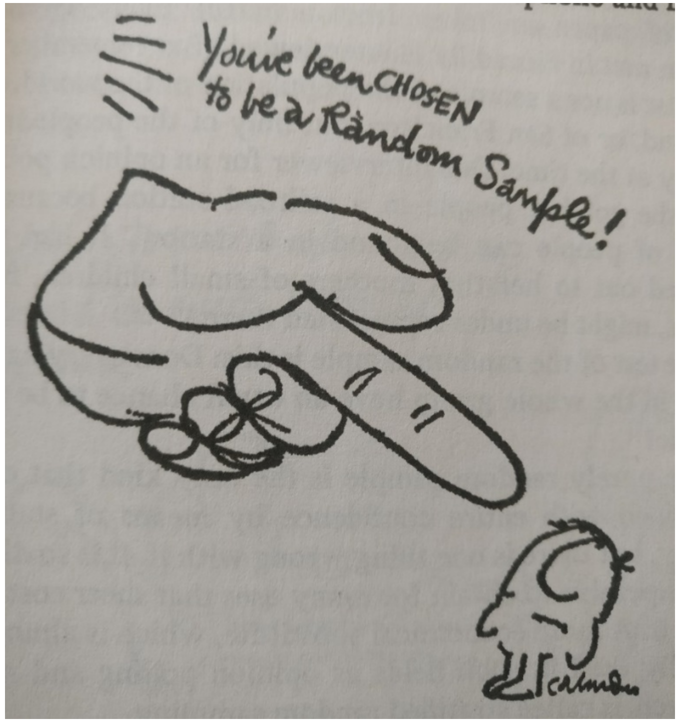
Pull balls from well mixed bins

Deal out shuffled cards, etc.

Use a computer program

Q: What computer program can we use?

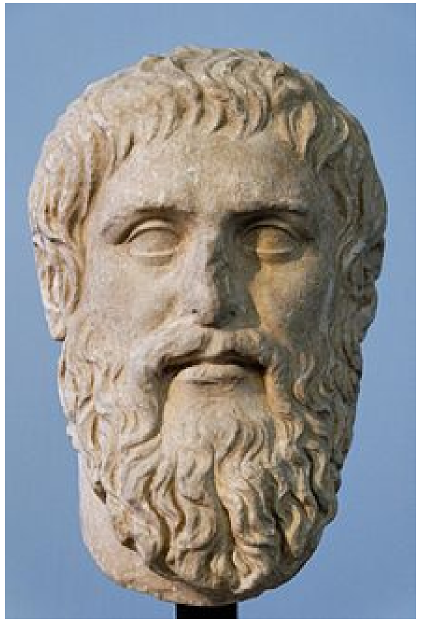
# Q: If you have questions about statistical bias where should you ask them?

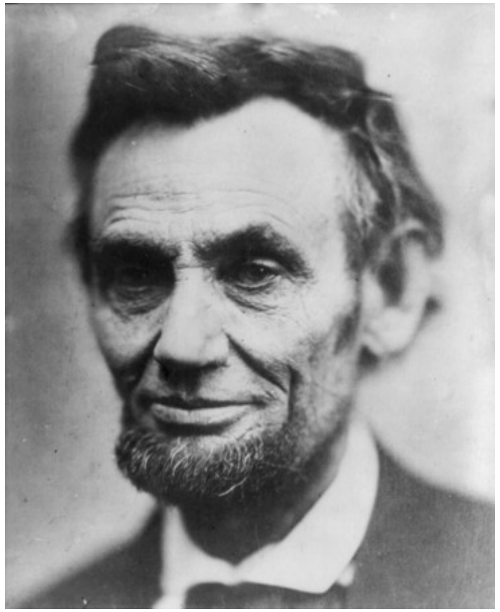


A: Ed discussions or come to class!

# From now on we are going to assume no bias!

Happy Plato and Lincoln



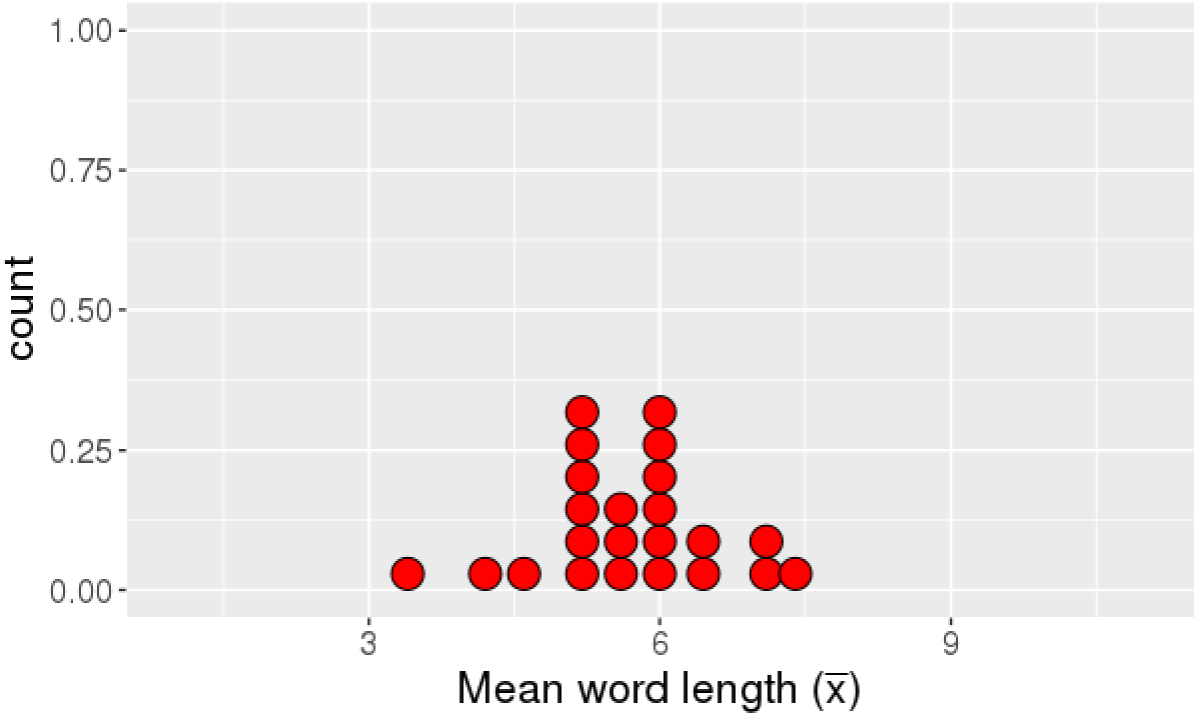


Our statistic values, on average, reflect the parameters

# Exploring sampling distributions in R

# For our distribution of Gettysburg word lengths…

The mean length of 10 words (x̅) from different student’s samples:



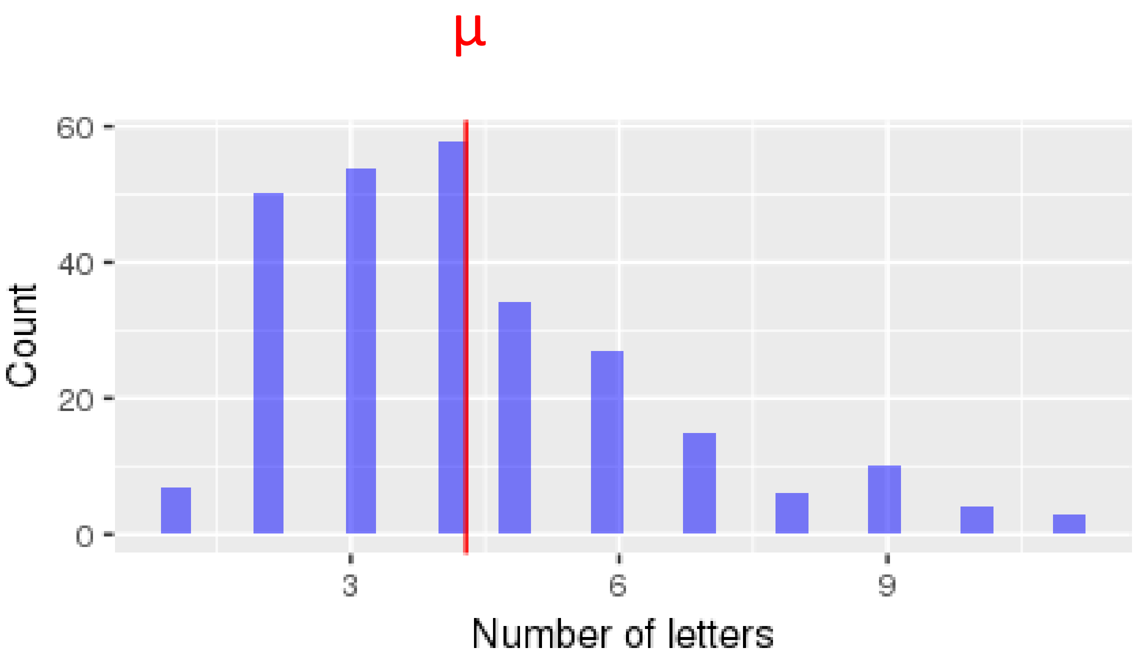
If we had infinite students taking samples yielding infinite statistics, we would have a **distribution** of statistics

# 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

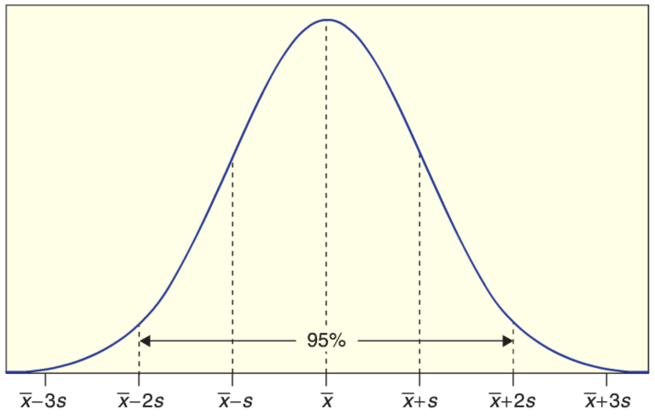
$\bar{x} = 5$

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!

# Let’s create a sampling distribution in R

Load the SDS100 library to make all SDS100 functions available

```
> library(SDS100)
```

Get the Gettysburg population data

```
> download_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

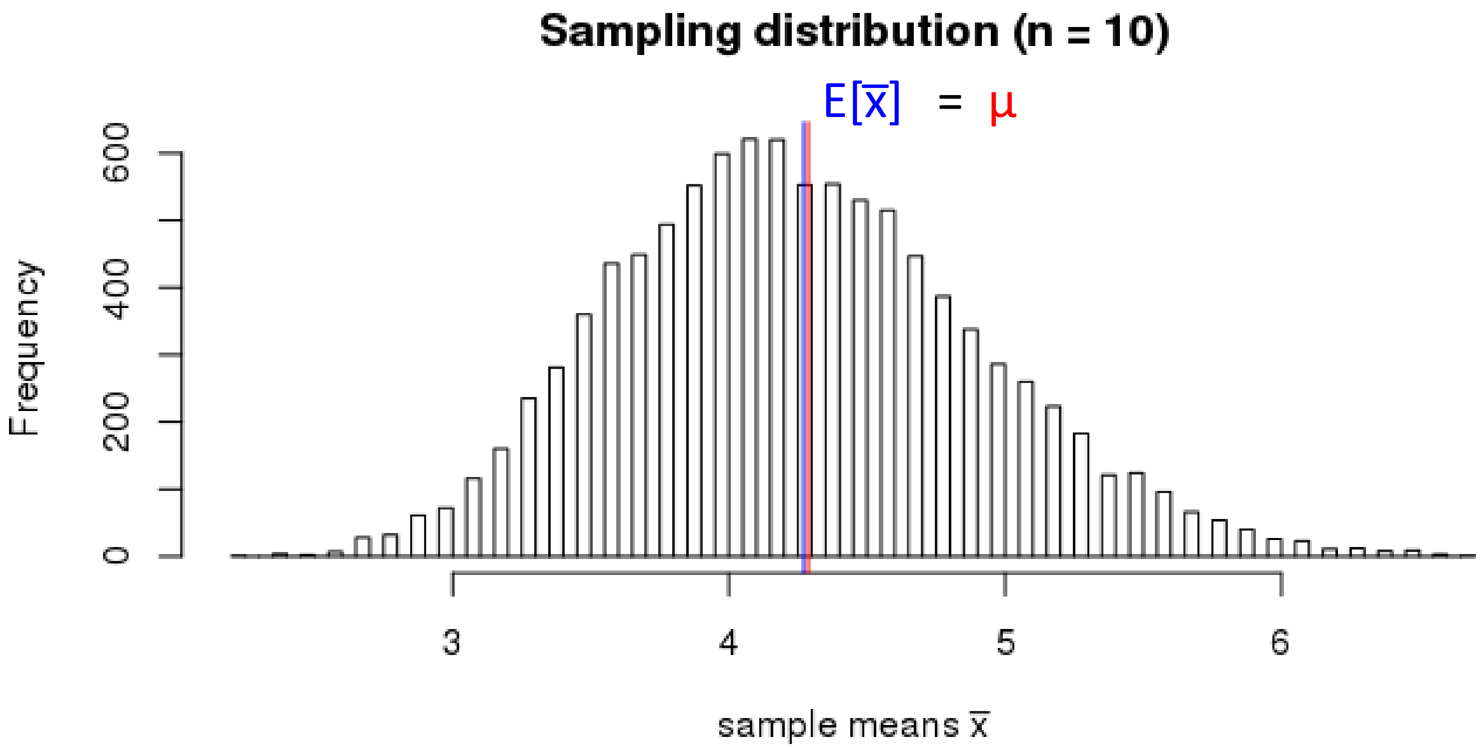
```
sampling_dist <- do_it(10000)  *  {

        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

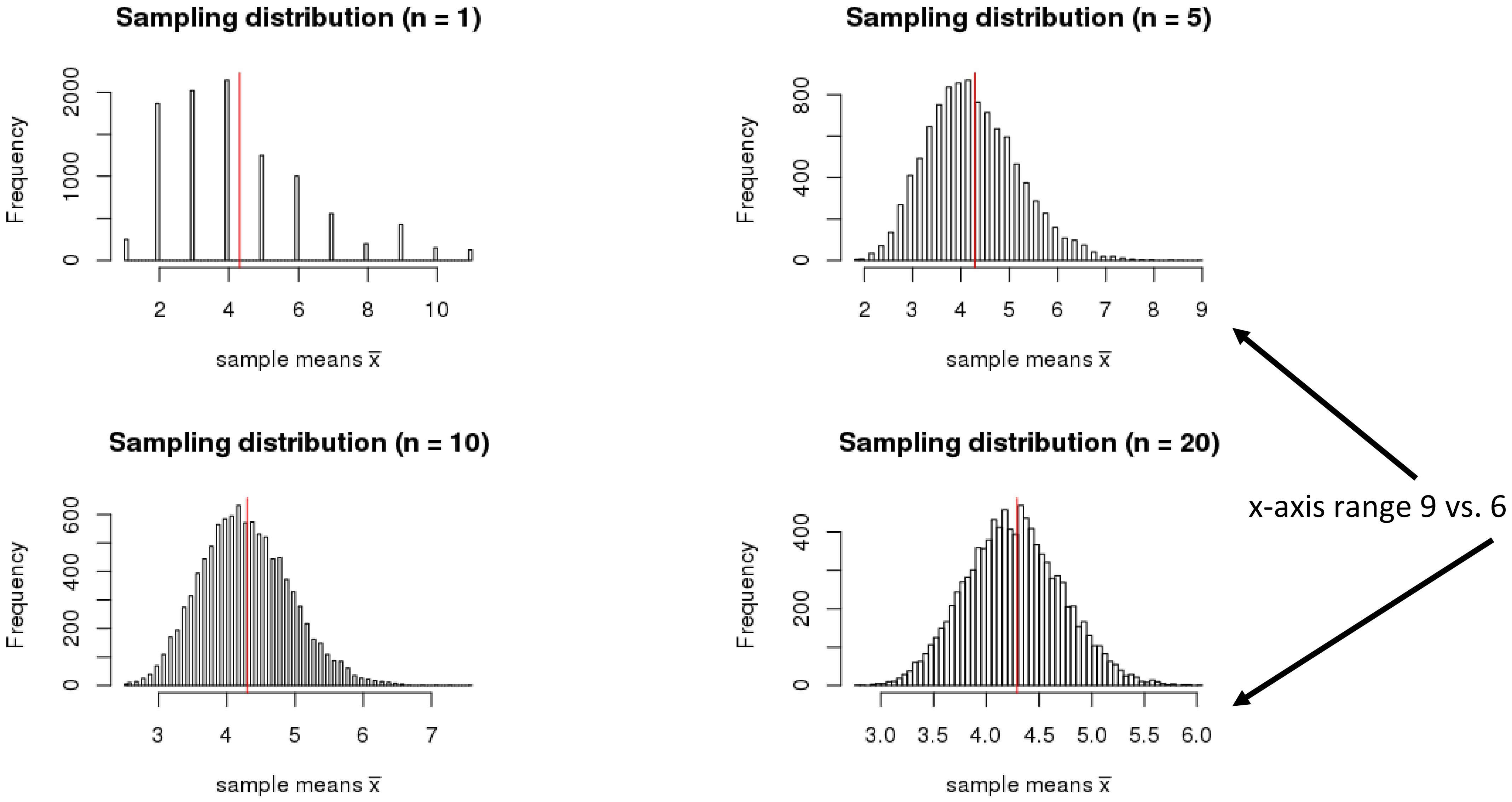
```
sampling_dist <- do_it(10000)  *  {

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

}


hist(sample_means, nclass = 100)
```

Gettysburg sampling distribution app



x-axis range 9 vs. 6

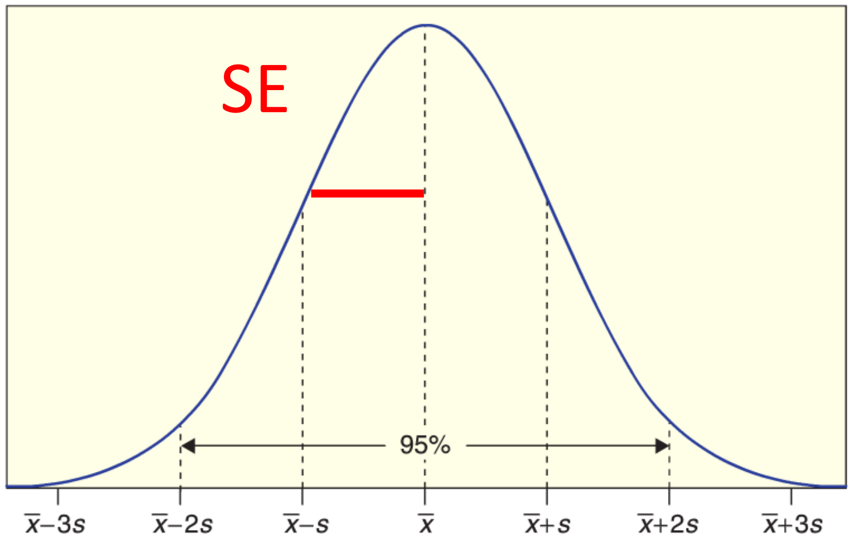
As the sample size n increases

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

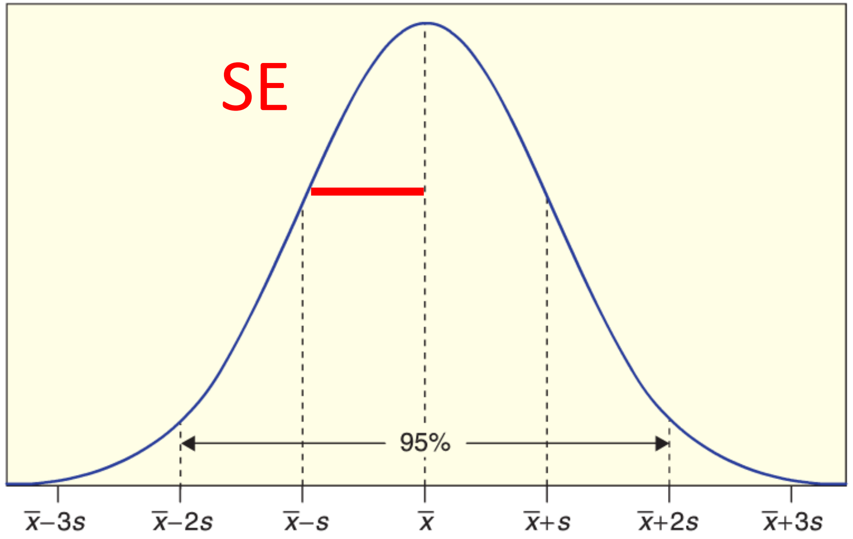
# The standard error

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

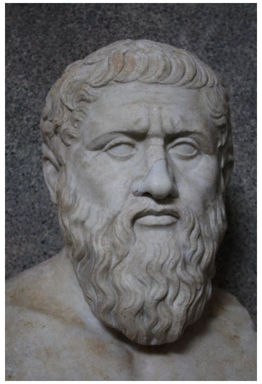
# Point estimates and confidence intervals

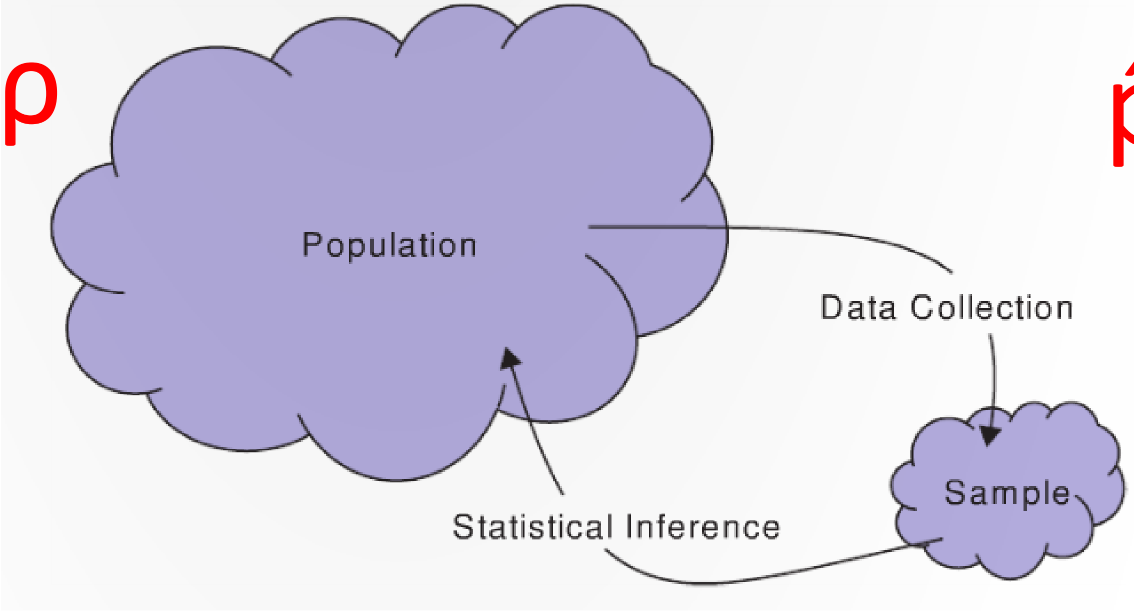
# Back to the big picture: Inference

**Statistical inference** is…?

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

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





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

# Point Estimate

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

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

54% of American approve of Biden’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: Gallup poll

54% of American approve of Biden’s job performance, plus or minus 3%

How do we interpret this?

Says that the population parameter ($\pi$) lies somewhere between 51% to 57%

i.e., if they sampled all voters the true population proportion ($\pi$) 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

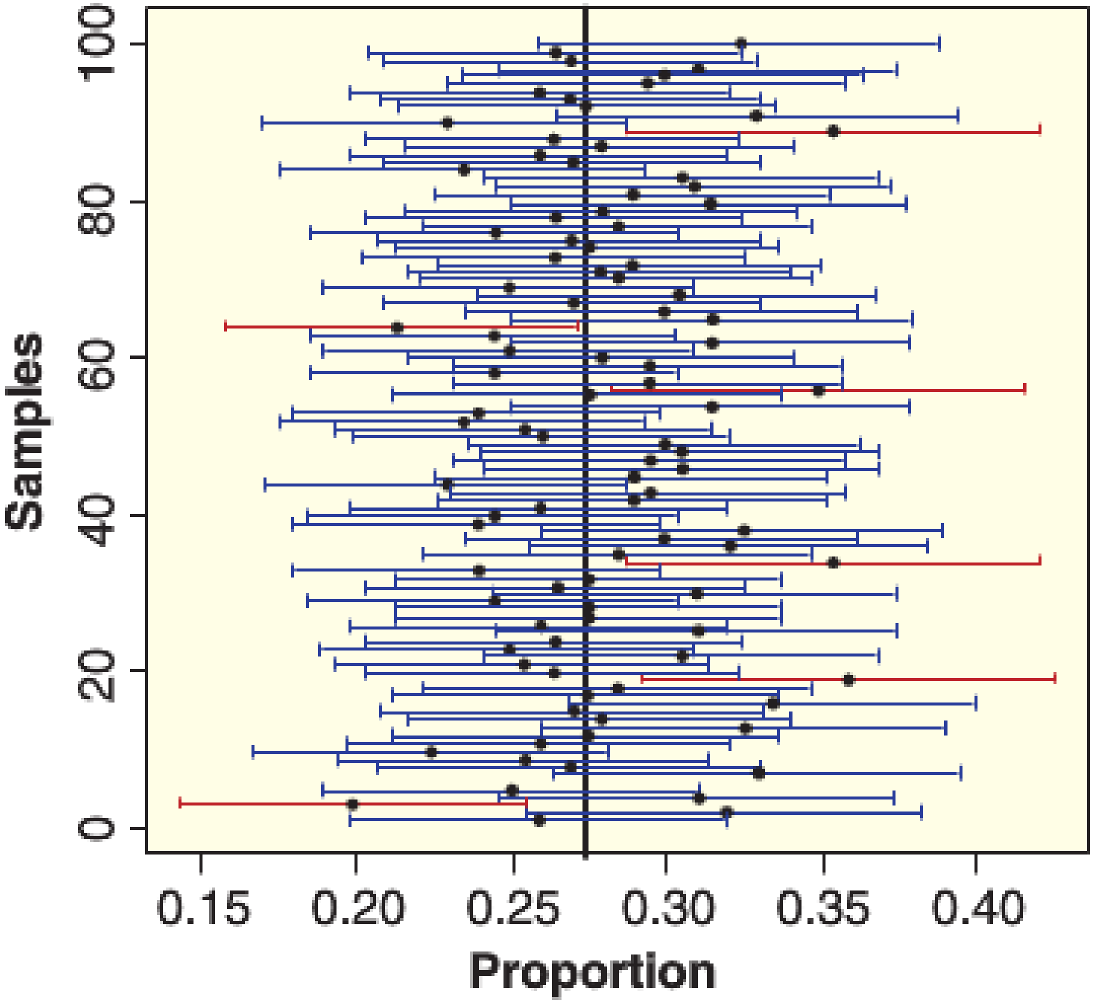
We toss intervals at it

95% of those intervals capture the parameter

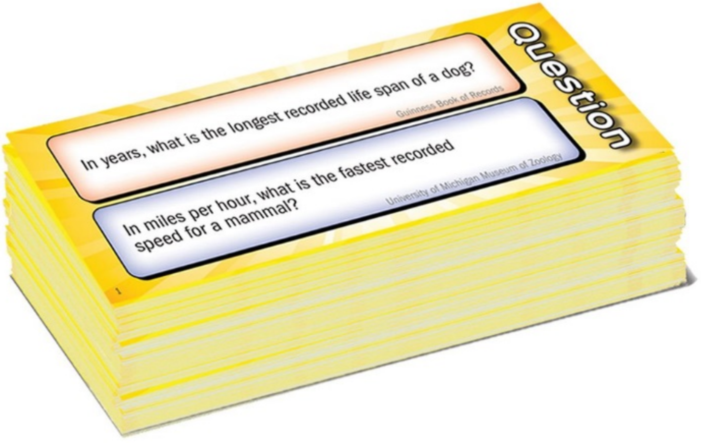
# Confidence Intervals

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

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



# Wits and Wagers:
# 90% confidence interval estimator


Question
In years, what is the longest recorded life span of a dog?
Guinness Book of Records
In miles per hour, what is the fastest recorded speed for a mammal?
University of Michigan Museum of Zoology
1

I will ask 10 questions that have numeric answers

Please come up with a range of values that contains the true value in it for 9 out of the 10 questions

- i.e., be a 90% confidence interval estimator

Enter your range of estimates for each question below as two numbers separated by a comma

- E.g., 10.2, 50.7

# Wits and Wagers…

**Question 1:** What percent of the world's surface is water?

**Question 2:** How many floors does the leaning tower of Pisa have?

**Question 3:** What year was the parking meter invented?

# Wits and Wagers…

**Question 4:** How many time zones does Russia have?

**Question 5:** In miles, how far does the average American drive each year?

**Question 6:** What percent of the world's population lives in the U.S.?

**Question 7:** On average, what percent of a watermelon's weight comes from water?

# Wits and Wagers…

**Question 8:** What percentage of Americans say that reading is their favorite leisure-time activity?

**Question 9:** In feet and inches, how tall was the tallest human in recorded history?

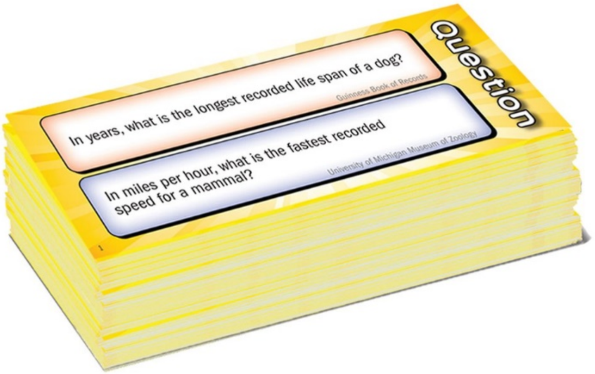
- Enter this as feet.inches e.g., 3.0, 5.11

**Question 10:** In what year was an ATM machine first installed in the U.S.?

# Please enter your interval on the Canvas quiz/survey

Enter your range of estimates for each question below as two numbers separated by a comma

- E.g., 10.2, 50.7



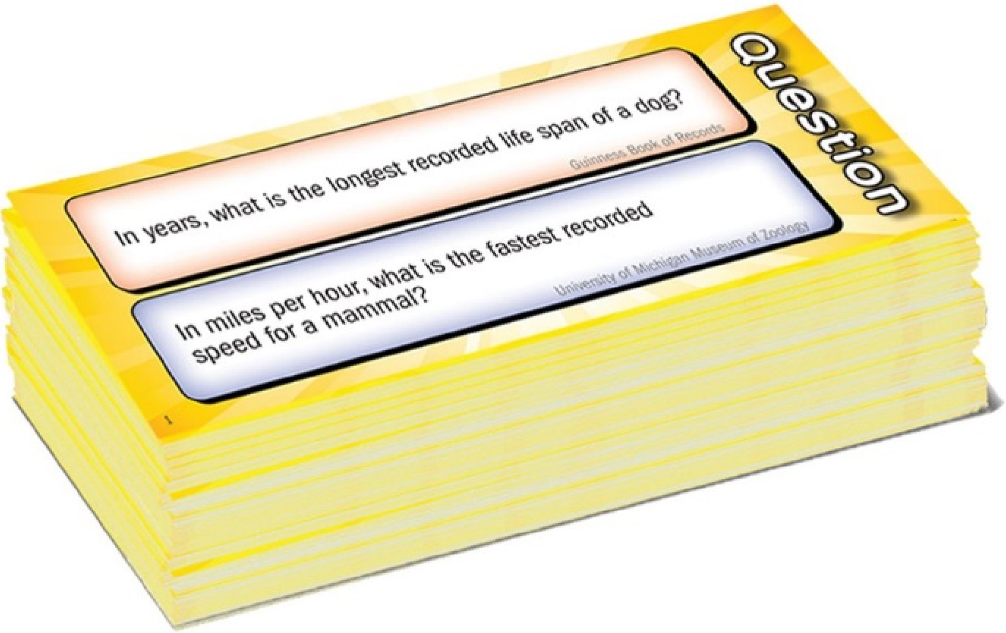
In years, what is the longest recorded life span of a dog?

Guinness Book of Records

In miles per hour, what is the fastest recorded speed for a mammal?

University of Michigan Museum of Zoology

# Answers…



In years, what is the longest recorded life span of a dog?

Guinness Book of Records

In miles per hour, what is the fastest recorded speed for a mammal?

University of Michigan Museum of Zoology

# Wits and Wagers…

**Question 1:** What percent of the world's surface is water?

- 71%

**Question 2:** How many floors does the leaning tower of Pisa have?

- 8

**Question 3:** What year was the parking meter invented?

- 1935

# Wits and Wagers…

**Question 4:** How many time zones does Russia have?

- 11

**Question 5:** In miles, how far does the average American drive each year?

- 13,476

**Question 6:** What percent of the world's population lives in the U.S.?

- 4.27%

**Question 7:** On average, what percent of a watermelon's weight comes from water?

- 92%

# Wits and Wagers…

**Question 8:** What percentage of Americans say that reading is their favorite leisure-time activity?

- 35%

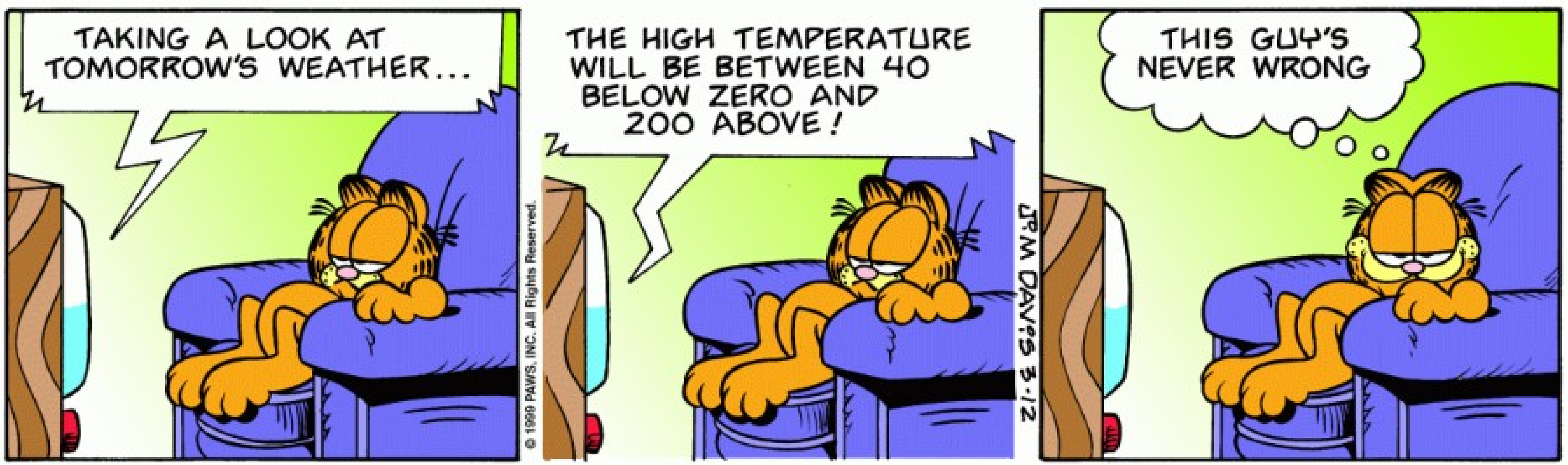
**Question 9:** In feet and inches, how tall was the tallest human in recorded history?

- 8’ 11”

**Question 10:** In what year was an ATM machine first installed in the U.S.?

- 1969

# 100% confidence intervals


TAKING A LOOK AT TOMORROW'S WEATHER...
THE HIGH TEMPERATURE WILL BE BETWEEN 40 BELOW ZERO AND 200 ABOVE!
THIS GUY'S NEVER WRONG
© 1990 PAWS, INC. All Rights Reserved.
JPM DAVIS 3-12

There is a <u>tradeoff</u> 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

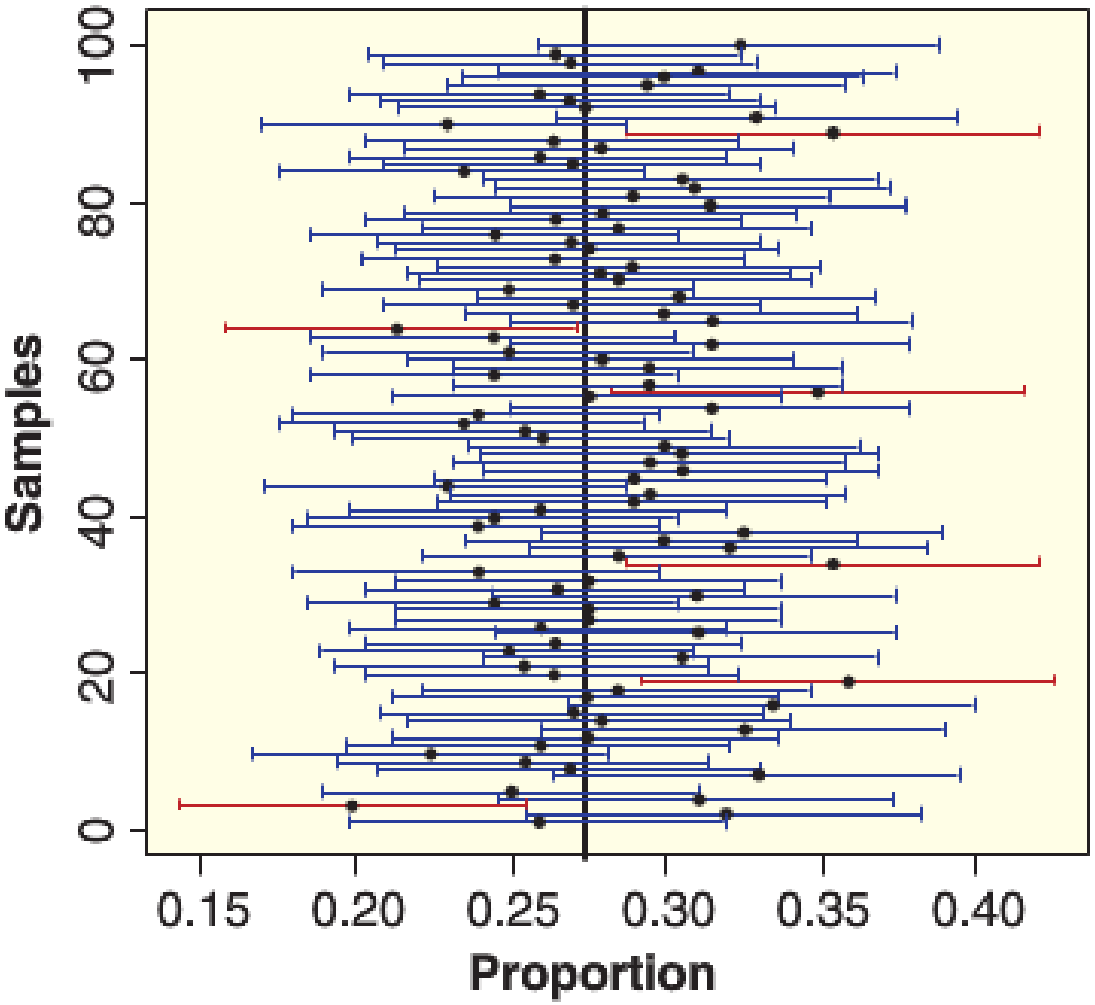
(for a 95% confidence interval)

# Confidence Intervals

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

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

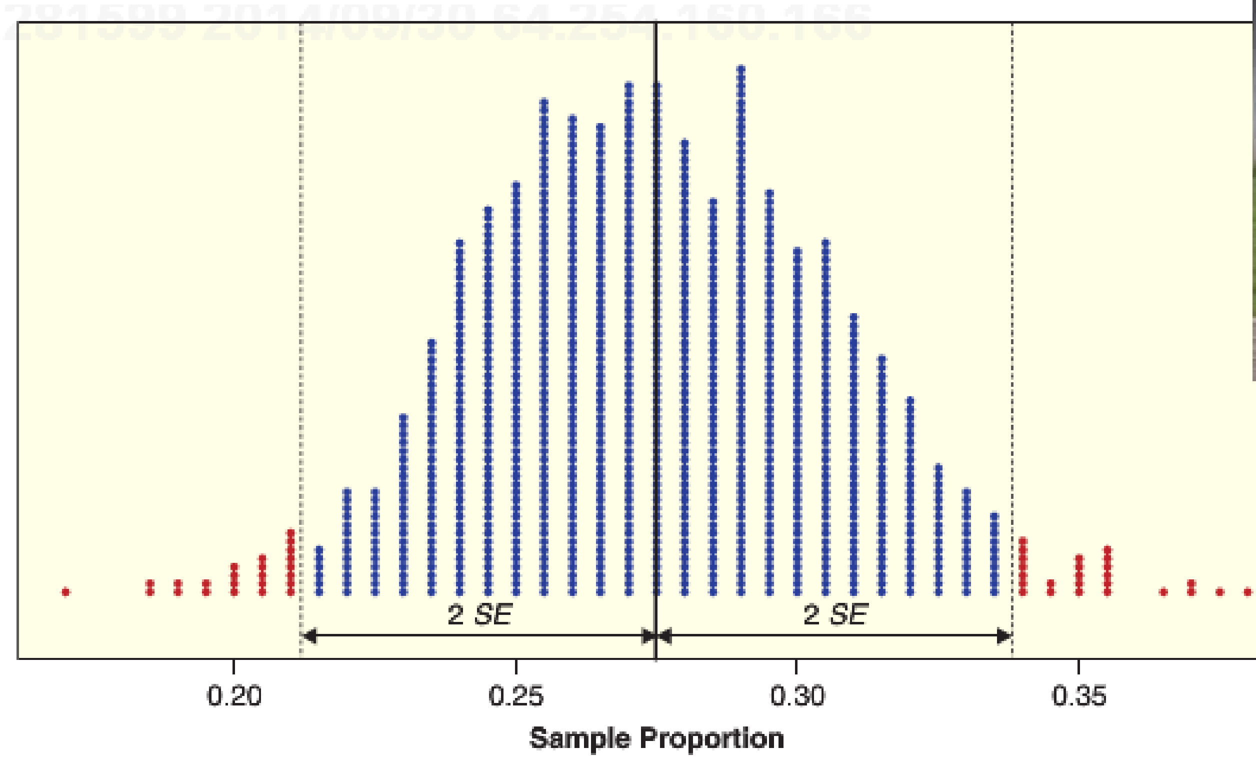
Right???



# Computing confidence intervals

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

# 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

# Next class

Using the bootstrap to estimate the SE…

# Homework 3

Homework 3 has been posted

- Due on Gradescope at 11:30pm on Sunday February 28th