Sampling distributions, standard errors, and confidence intervals

Overview

Review and continuation of sampling and bias

Sampling distributions

Exploring sampling distributions in R and the Standard Error

If there is time

- Point estimates and confidence intervals

Announcement

Homework 4 has been posted!

It is due on Gradescope on Sunday September 28th at 11pm

- **Be sure to mark each question on Gradescope!**

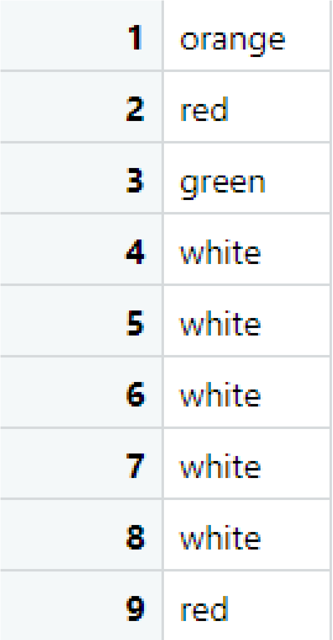
My office hours will be ending a little early on Wednesday

- 2pm-2:45pm

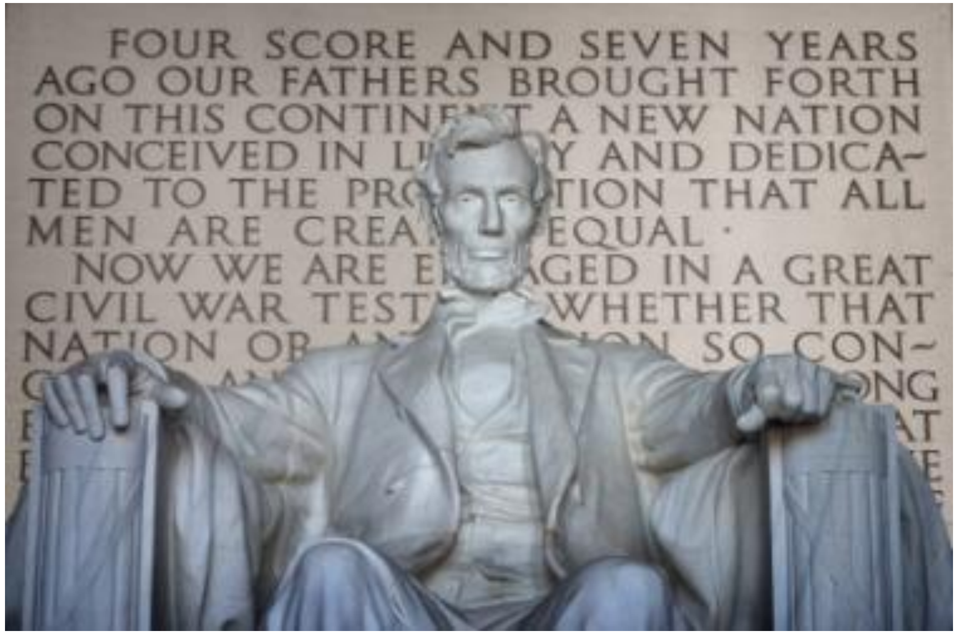
The material this week is going to be a bit more conceptually challenging.
Please attend the practice sessions to reinforce your understanding!

Review: sampling and sampling distributions

Review: sampling

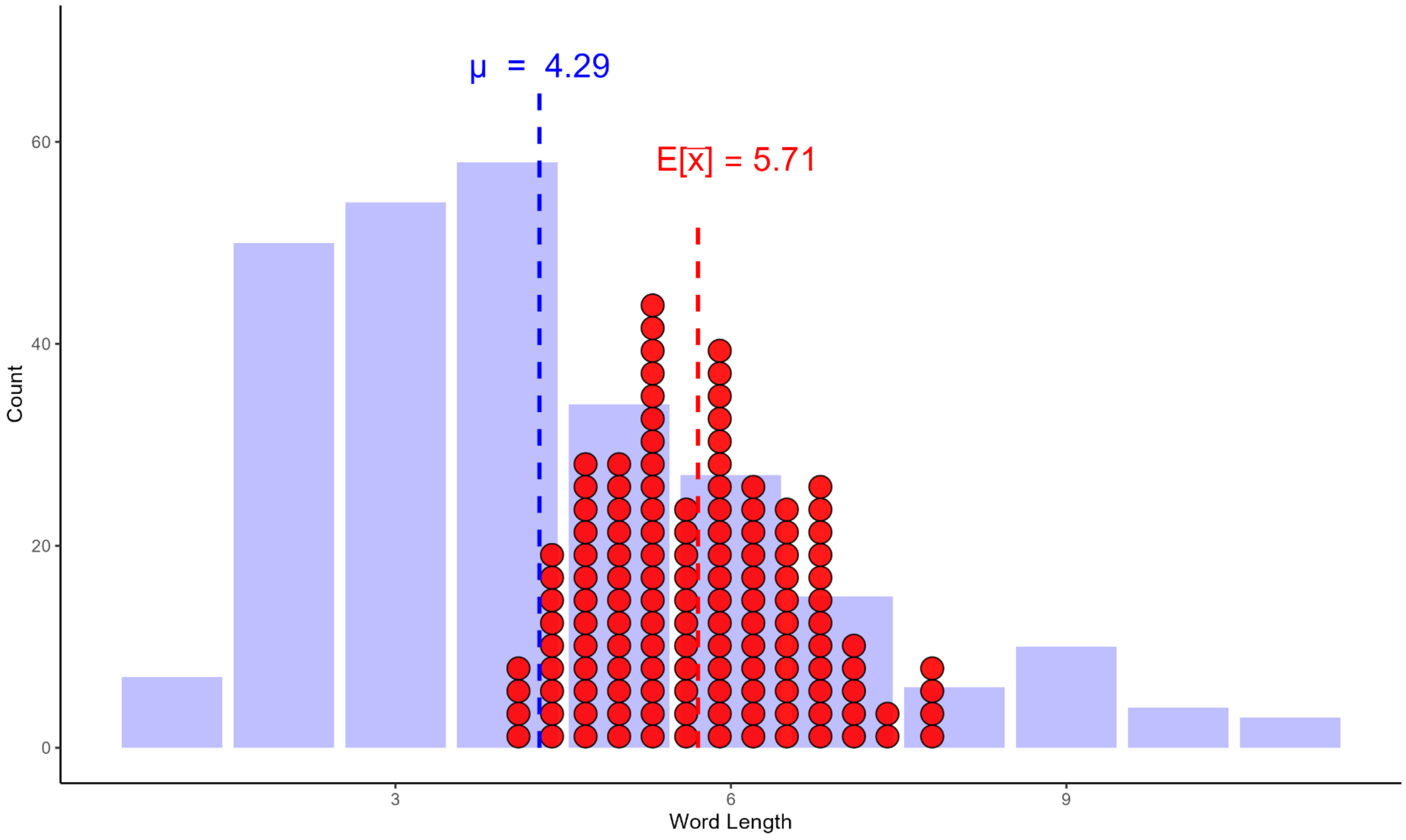


1	orange
2	red
3	green
4	white
5	white
6	white
7	white
8	white
9	red


FOUR SCORE AND SEVEN YEARS
AGO OUR FATHERS BROUGHT FORTH
ON THIS CONTINENT A NEW NATION
CONCEIVED IN LI Y AND DEDICA-
TED TO THE PRO 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

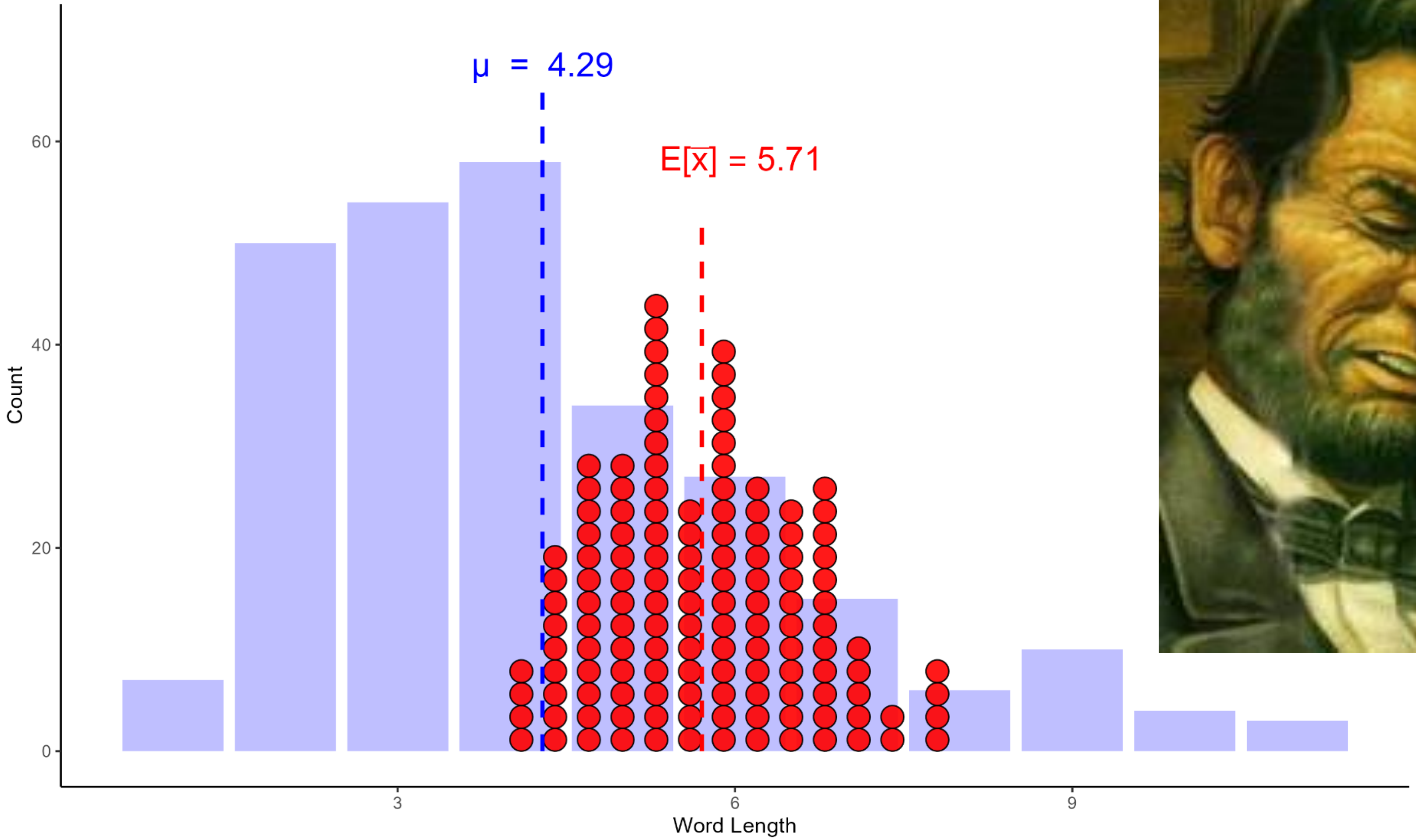


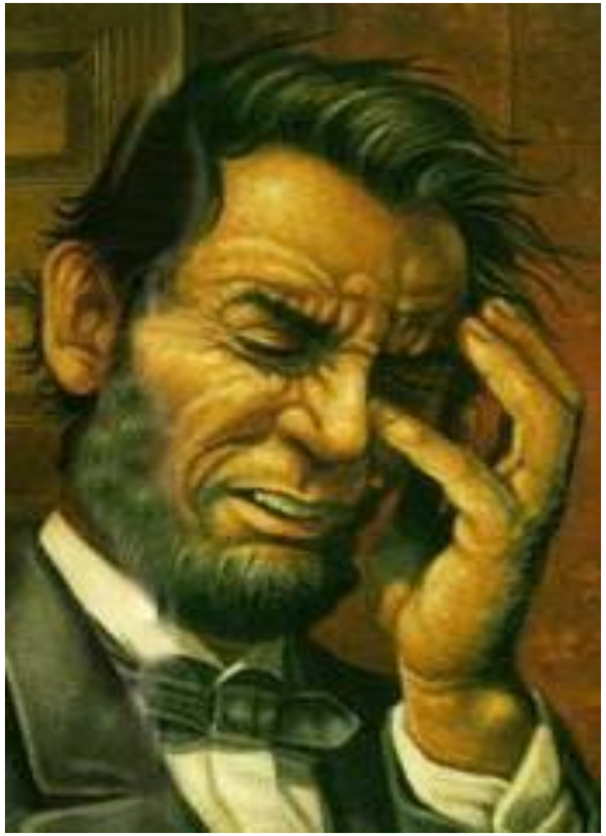
Bias and the Gettysburg address word length distribution

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

Here:

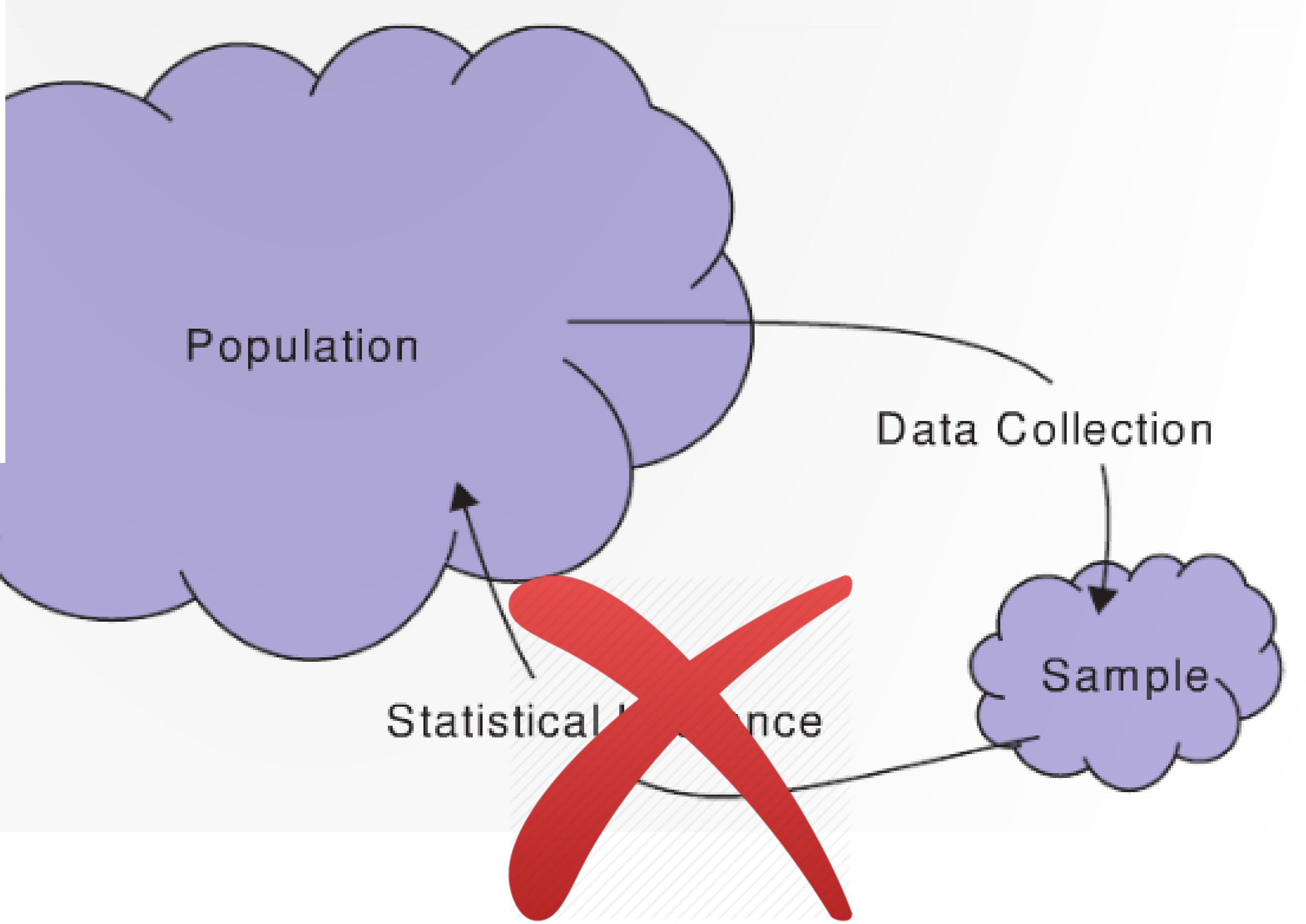
$E_s[\overline{x}] \neq \mu$





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?

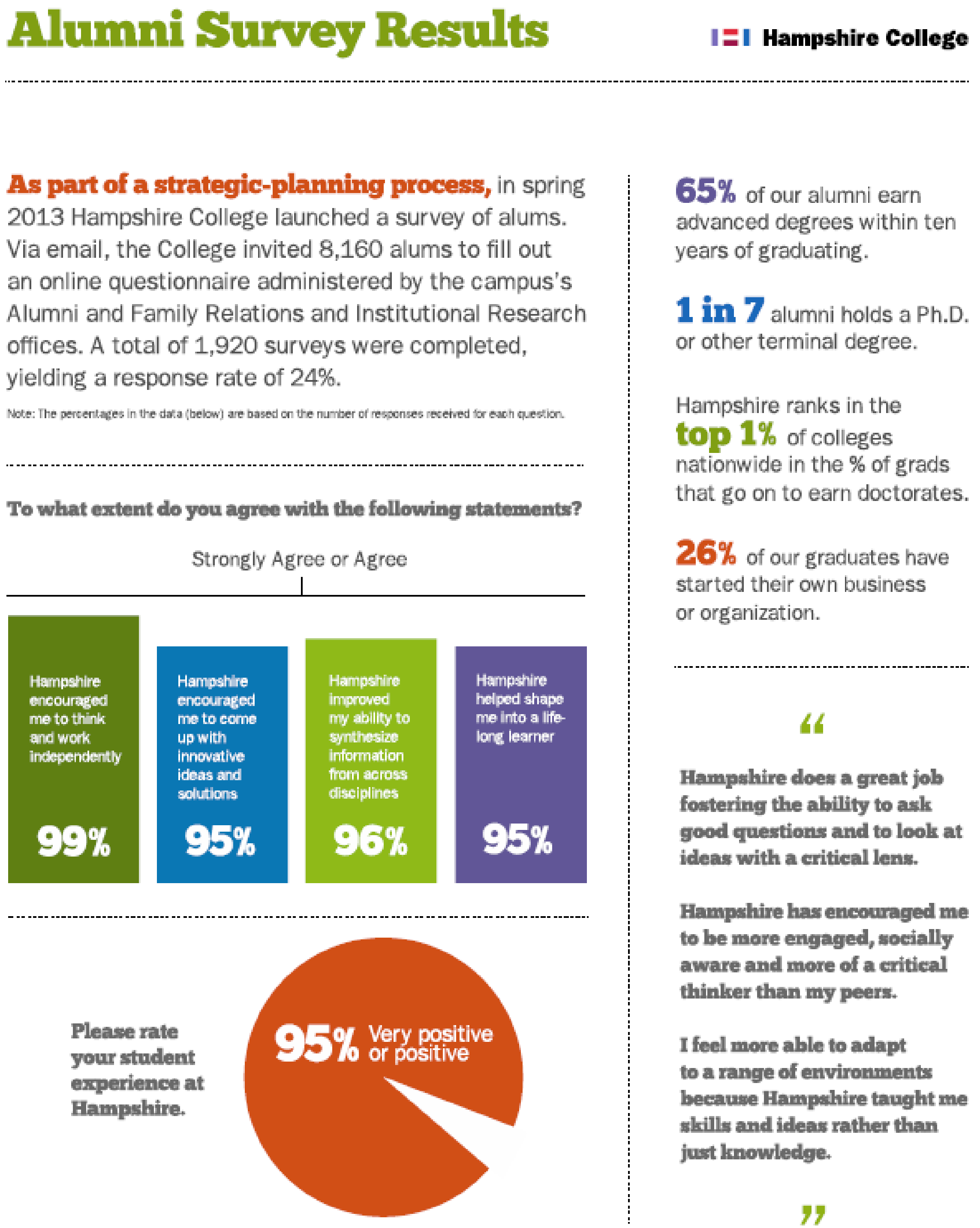
1948 US election: Dewey vs. Truman

Suppose there was a poll for the Truman/Dewey election that had randomly chosen 6,000 people from all voters 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

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.**
>
> ”

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 people left-handed in the US, randomly sampling Yale students might be good enough

Need to think carefully to avoid bias!

Statistics requires thought!

Use your own reasoning:

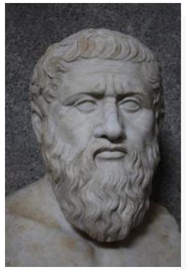
- What is the population I am interested in?
- Does the sample reflect the population of interest?
- Be your own worst critic!

Q: How can we prevent sampling bias?

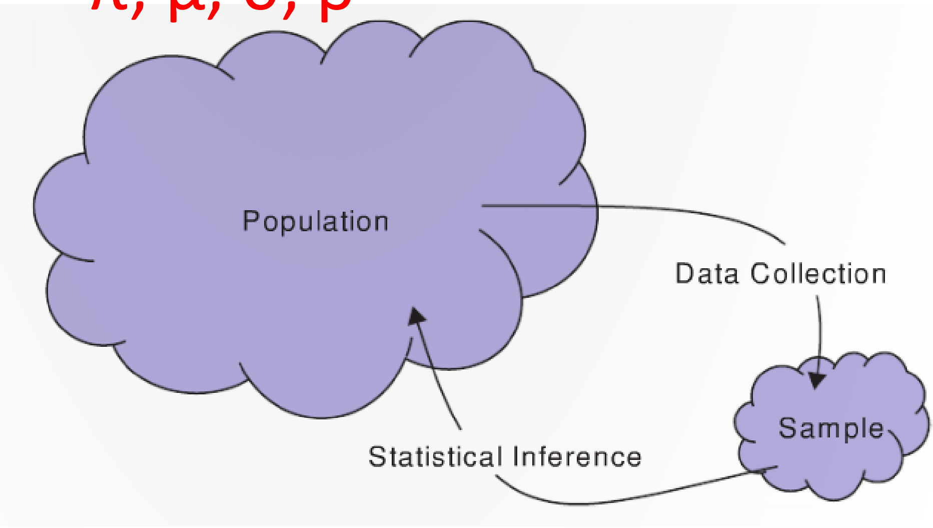
To prevent bias, use a **simple random sample**

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

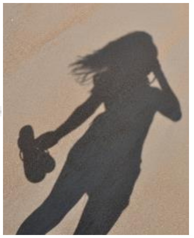
This allows for generalizations to the population!



π, μ, σ, ρ



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



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 a computer program

Questions about statistical bias?

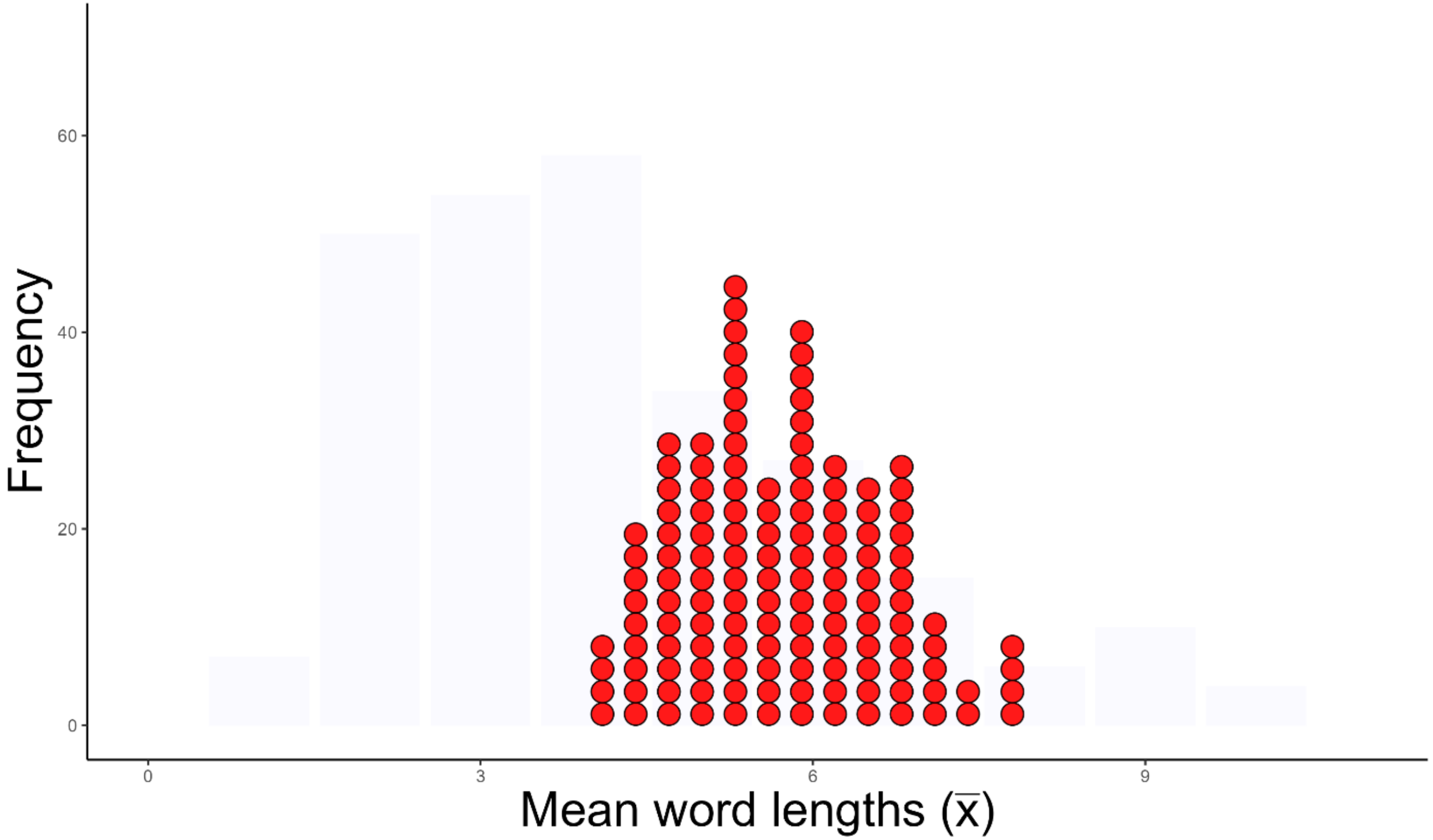
From now on we are going to assume no bias!

Our statistic values, on average, reflect the parameters

Sampling distributions

Recall for our distribution of Gettysburg word lengths…

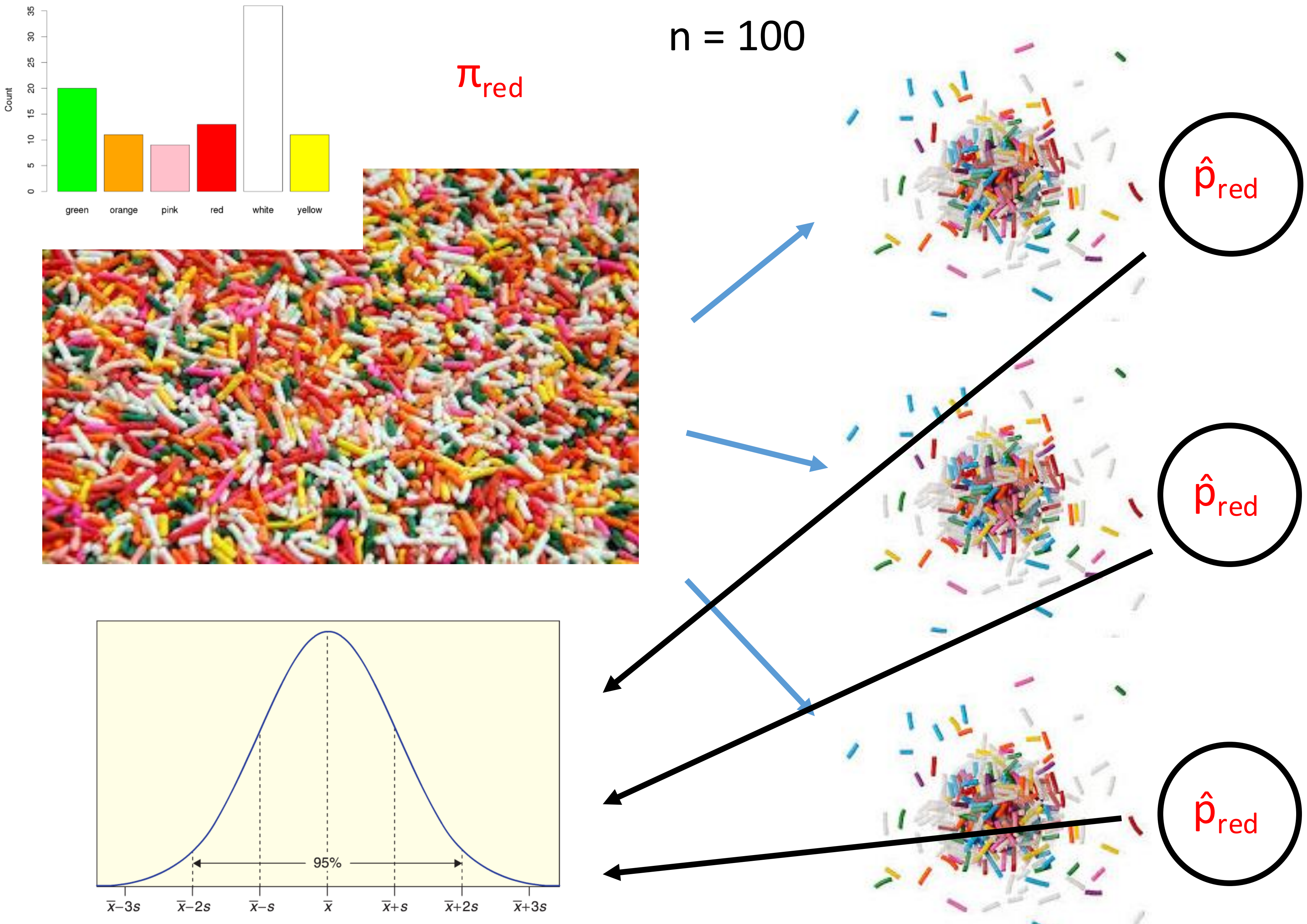
Q: What does each dot that is plotted correspond to?



Sampling distribution

A **sampling distribution** is the distribution of sample statistics 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



π_{red}

$n = 100$

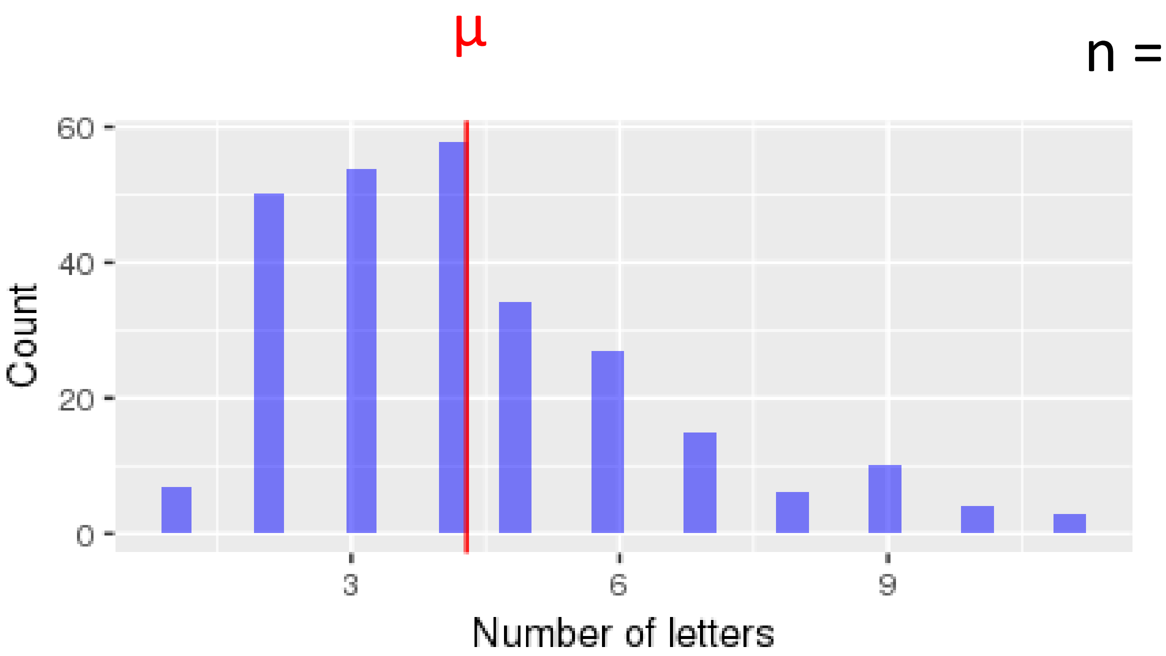
$\hat{p}_{red}$

$\hat{p}_{red}$

$\hat{p}_{red}$

Sampling distribution!

Gettysburg address word length sampling distribution



n = 10

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

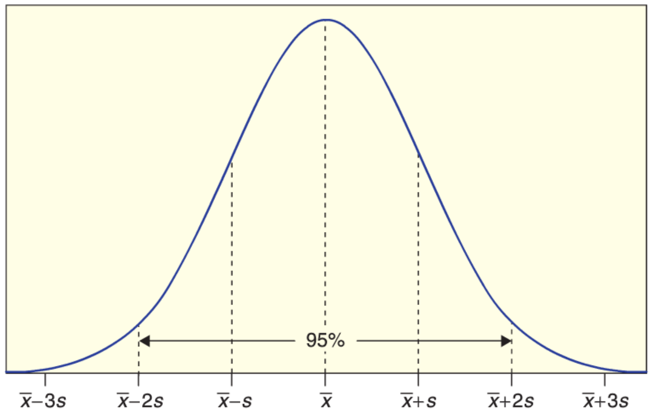
$\bar{x} = 5$

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 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

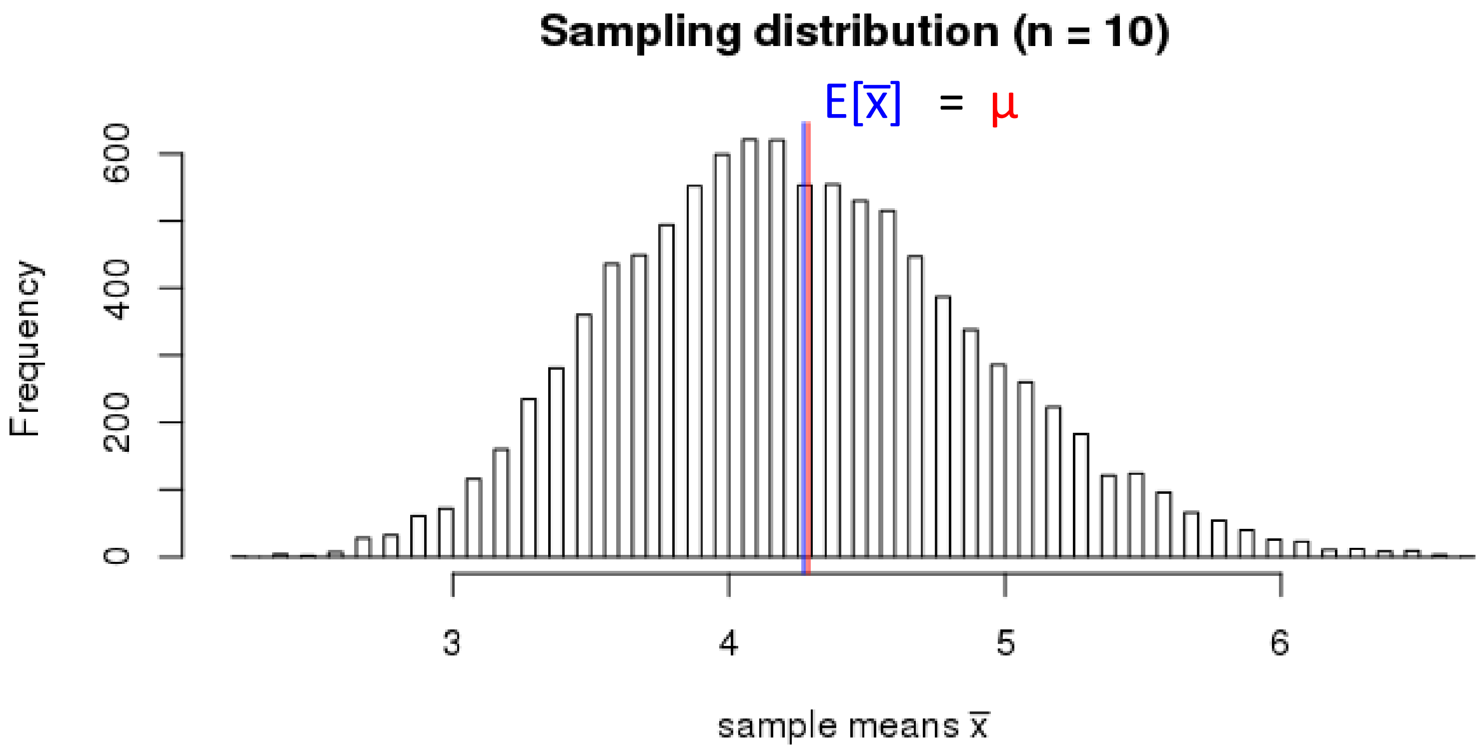
```
do_it(100)  *  {

        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)
```

Let’s try it ourselves in R!

Sampling distribution in R



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

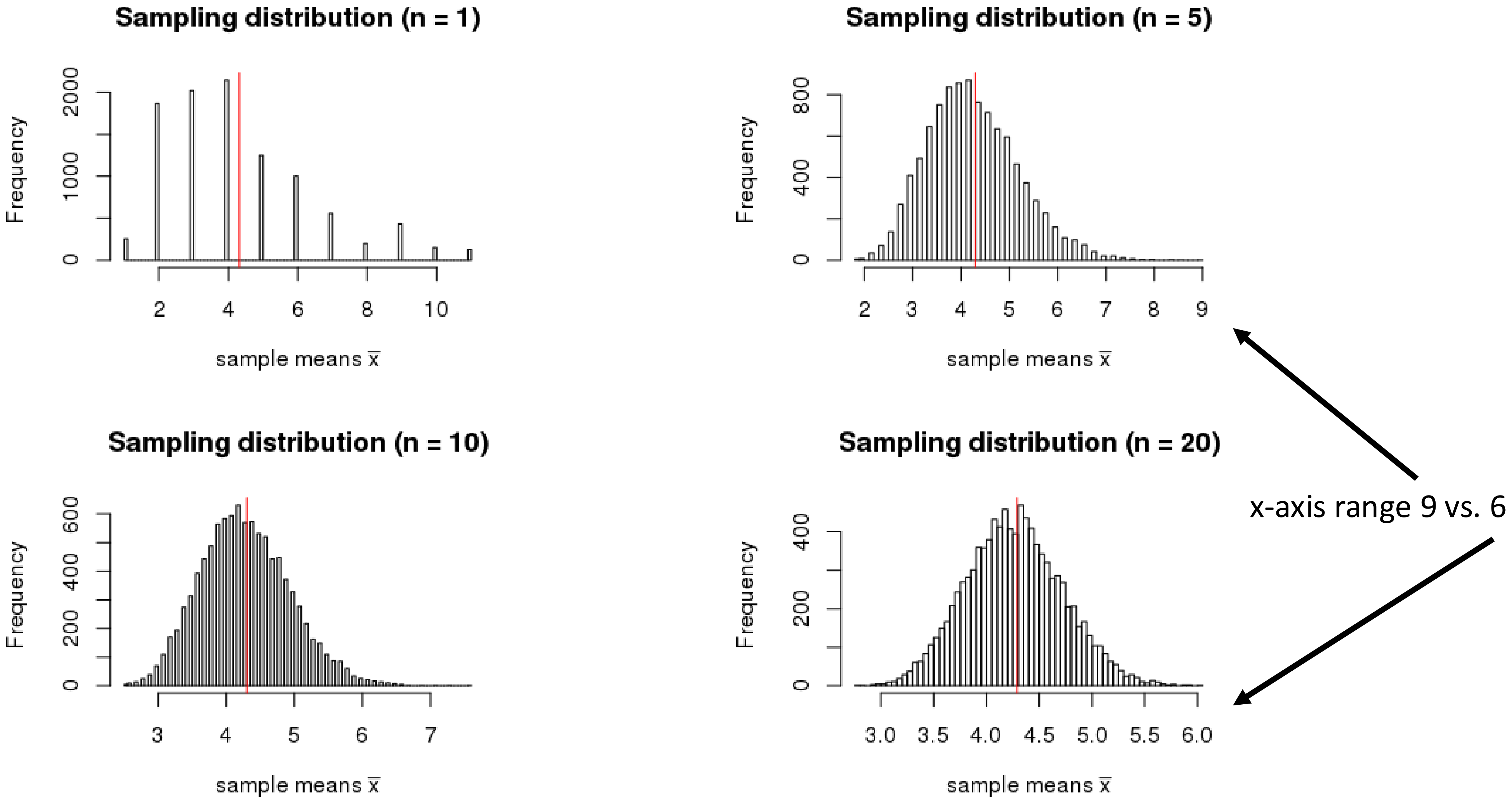
Changing the sample size n

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



Sampling distribution (n = 1)

Sampling distribution (n = 5)

Sampling distribution (n = 10)

Sampling distribution (n = 20)

x-axis range 9 vs. 6

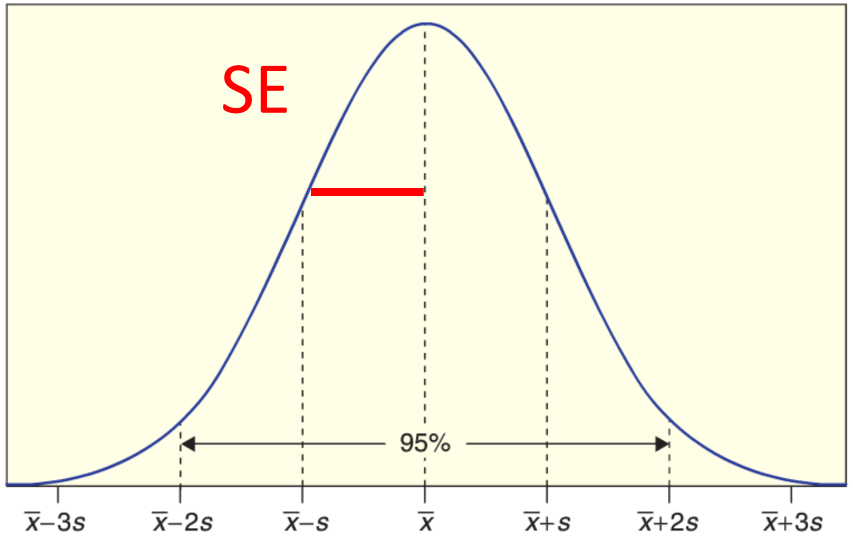
As the sample size n increases

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

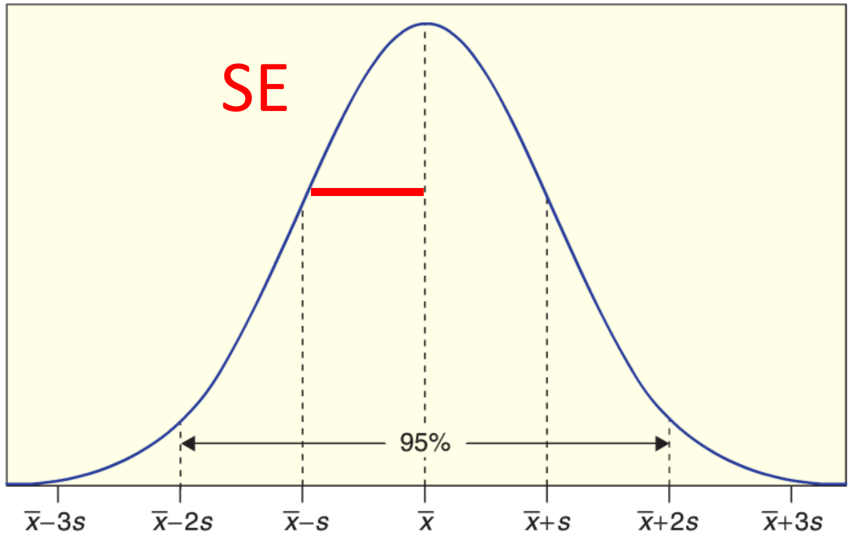
The standard error

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?

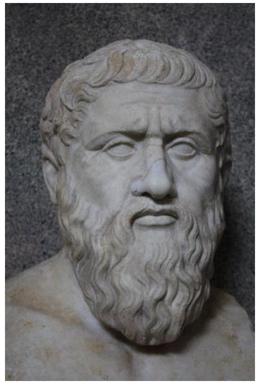
Point estimates and confidence intervals

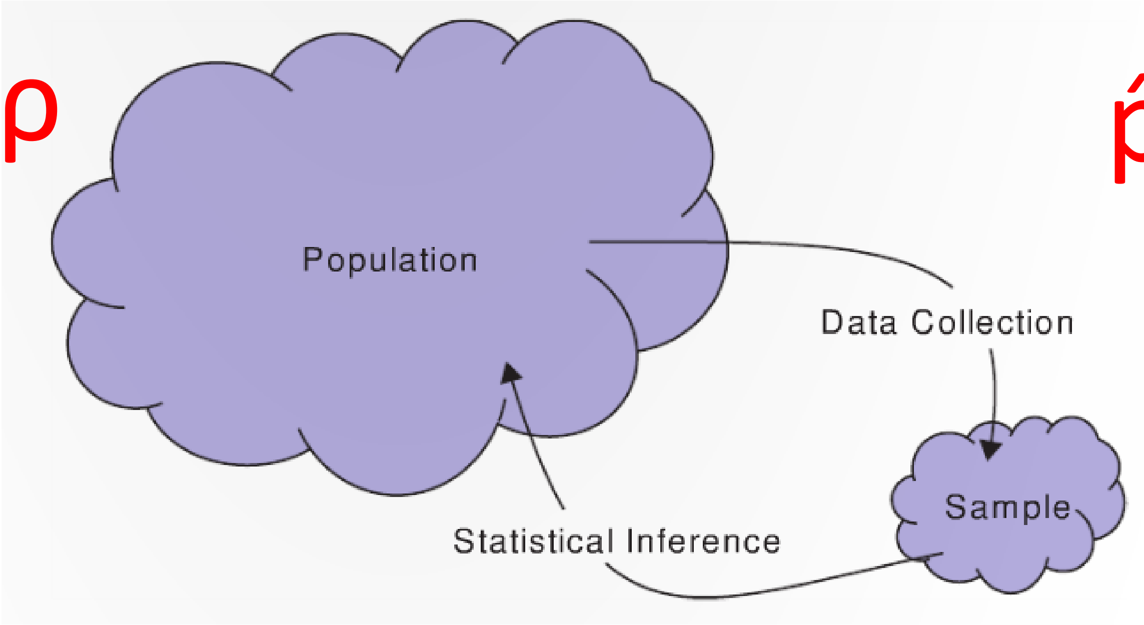
Back to the big picture: Inference

Statistical inference is…?

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

π, μ, σ, ρ





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

Point Estimate

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

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

40% of American approve of Trump's job performance according to a recent Gallup poll

Q: What are π and $\hat{p}$ here?

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

40% 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 37% to 43%

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

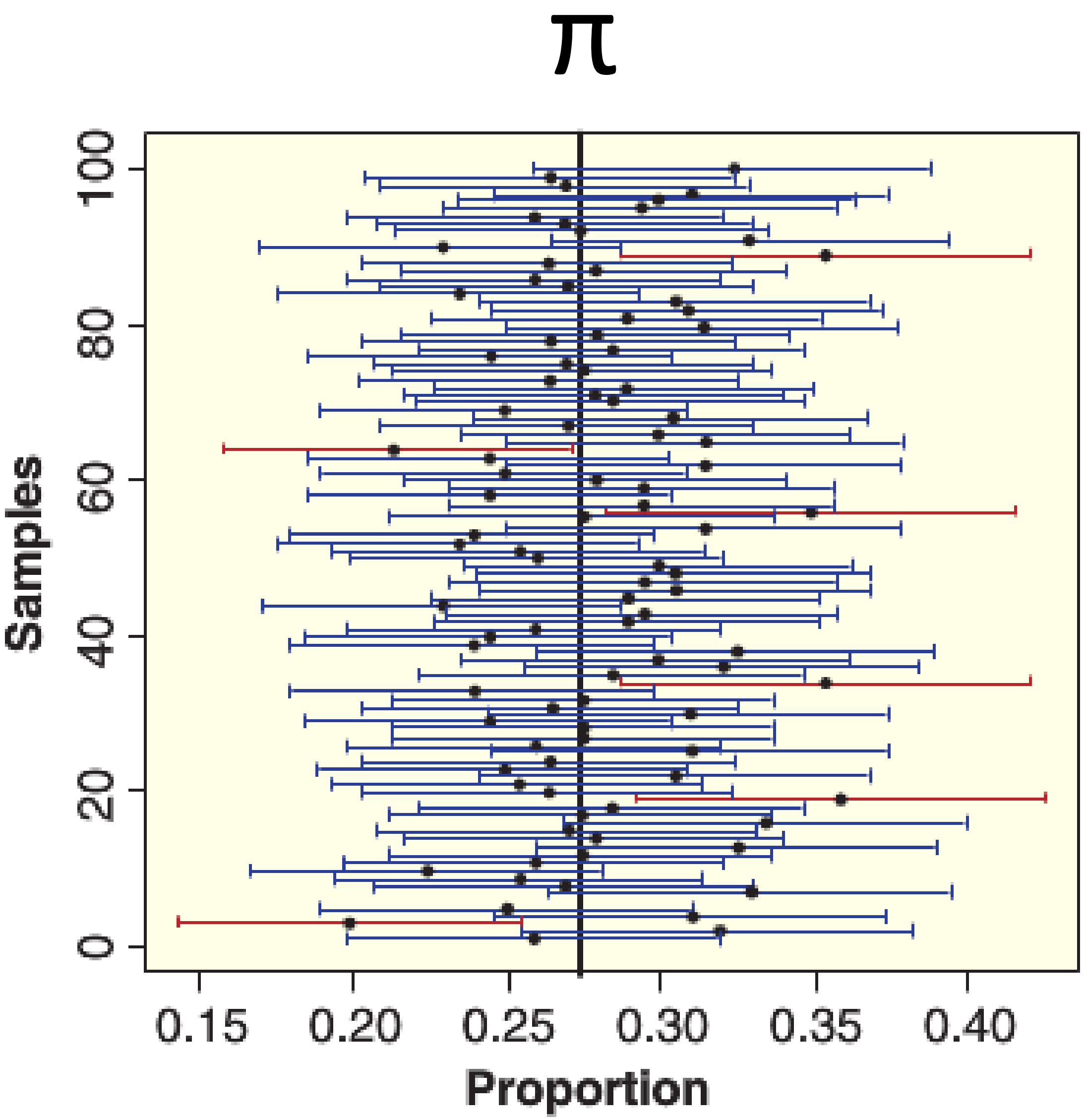
We toss intervals at it

95% of those intervals capture the parameter

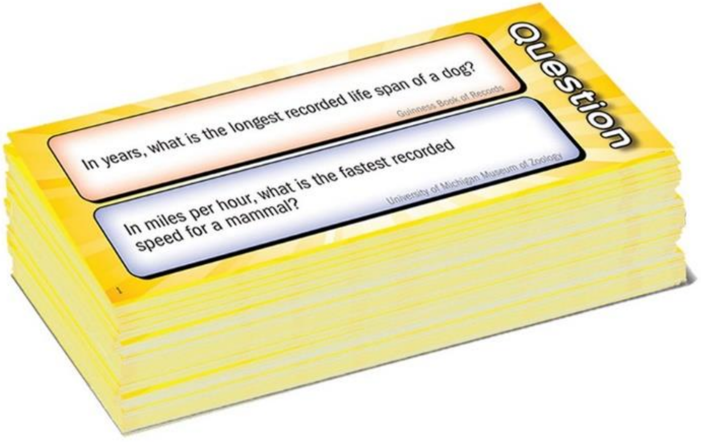
Confidence Intervals

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

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



Wits and Wagers:
90% confidence interval estimator



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

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

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

Write your range of estimates for each question. as two numbers

- E.g.,   [10.2   to   50.7]

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)