Sampling distributions,
standard errors, and confidence intervals

Overview

Quick review of bias and sampling distributions

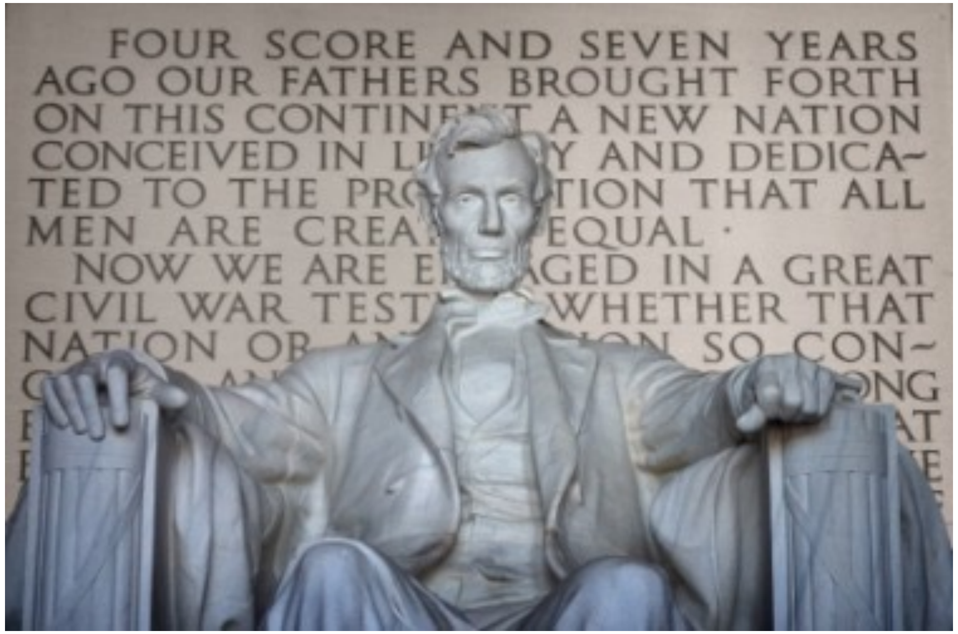
Exploring sampling distributions in R and the Standard Error

Point estimates and confidence intervals

Review: sampling and sampling distributions

Review: sampling

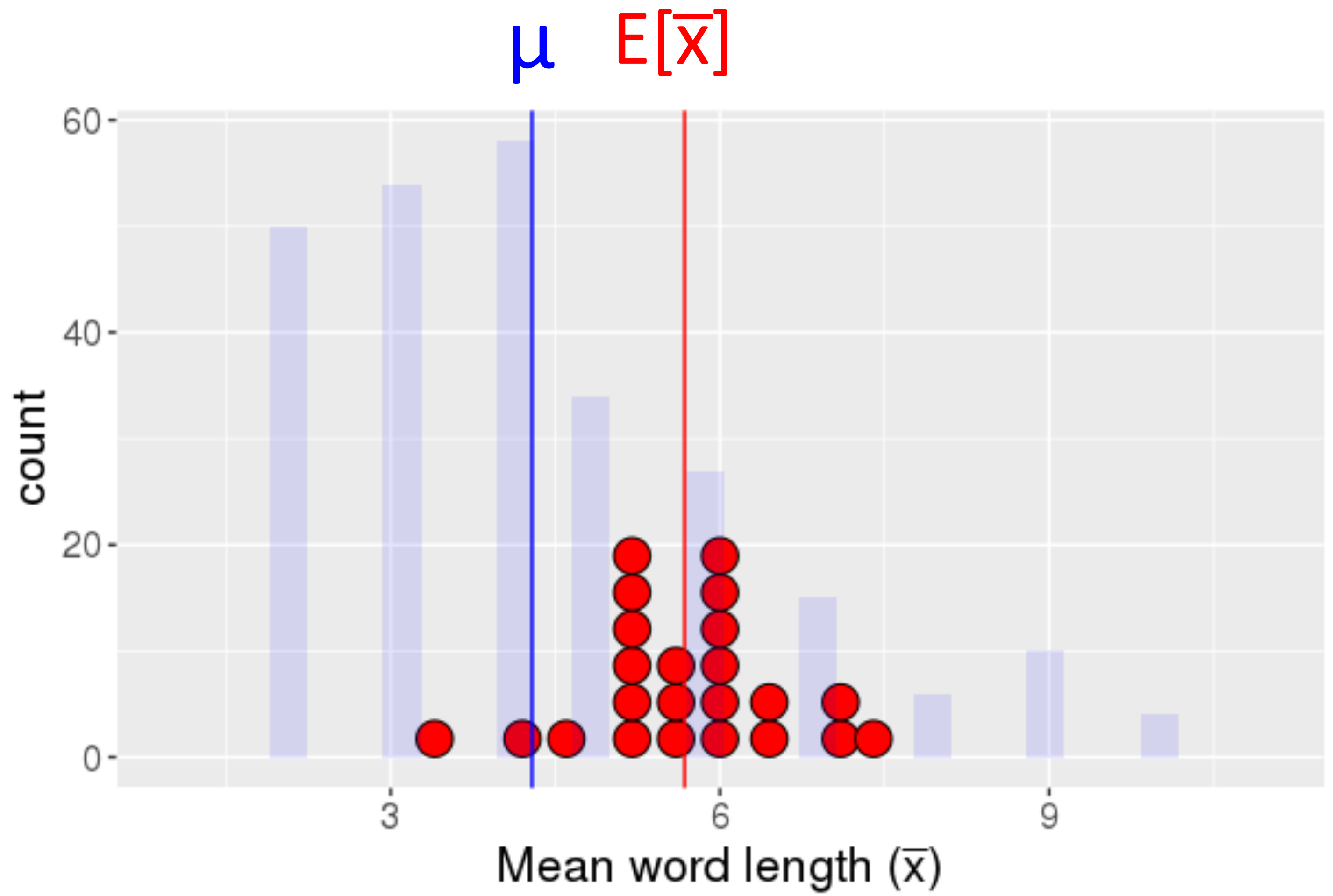
1	orange
2	red
3	green
4	white
5	white
6	white
7	white
8	white
9	red



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

A: ***n***

Bias and the Gettysburg address word length distribution

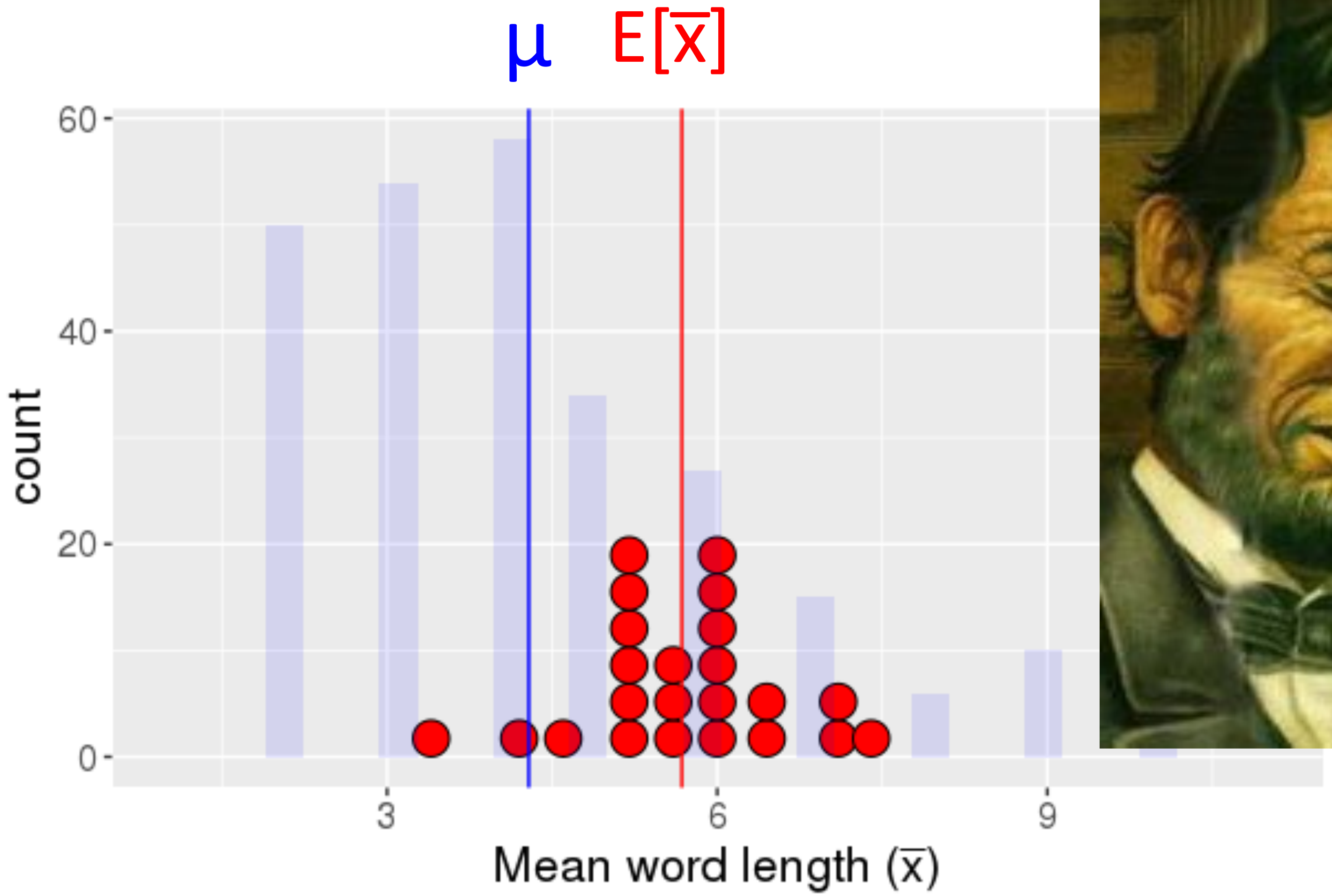


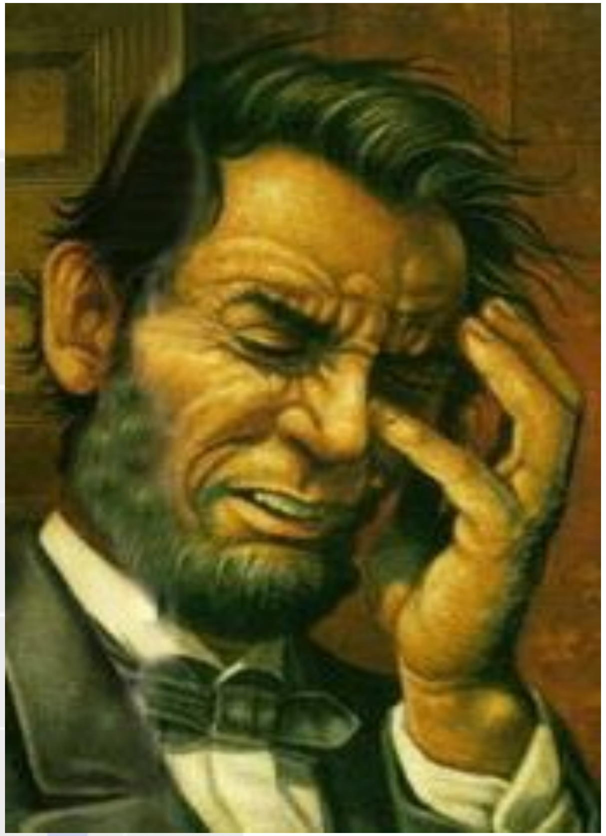
Bias and the Gettysburg address word length distribution

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

Here:

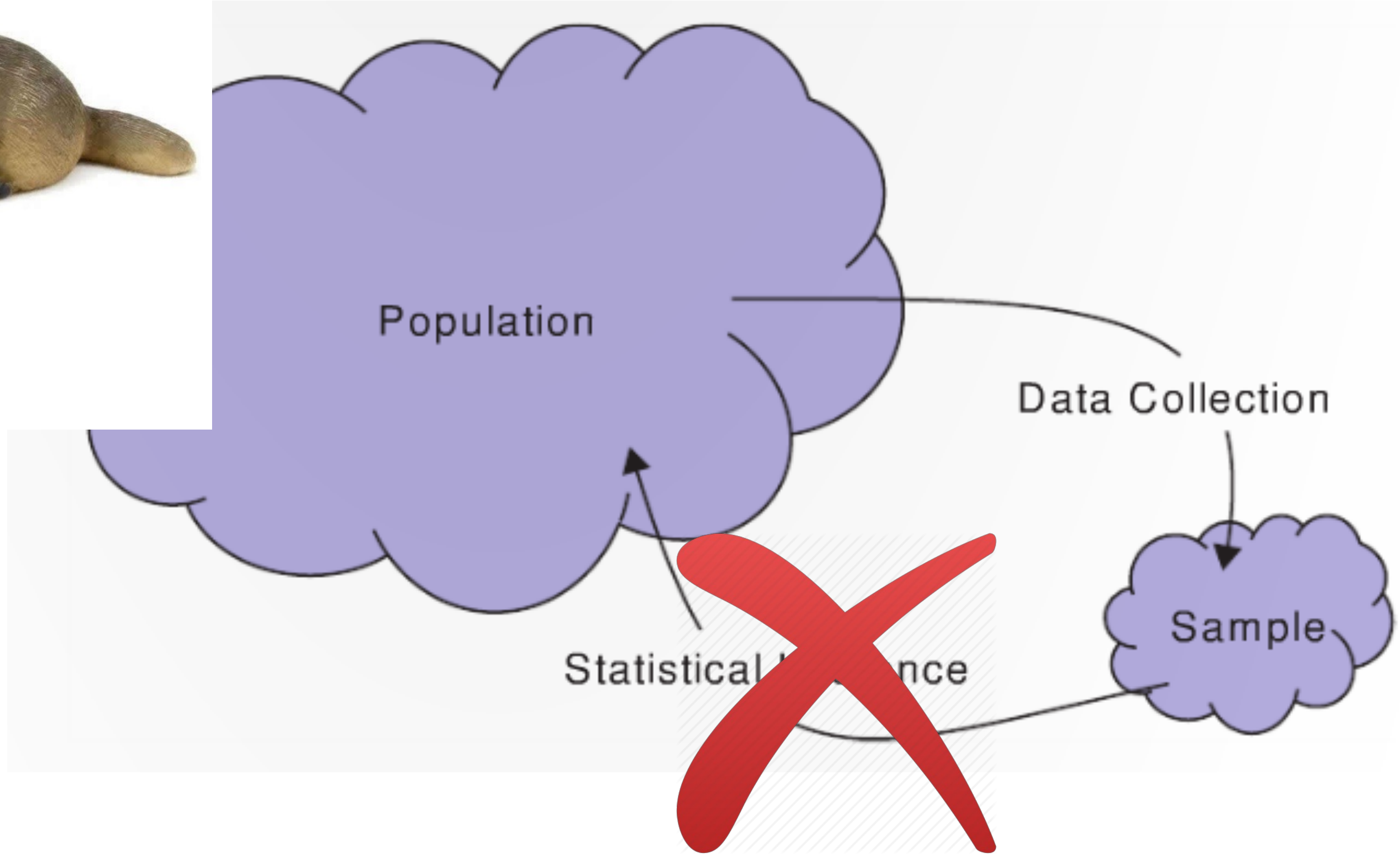
$E[\bar{x}] \neq \mu$



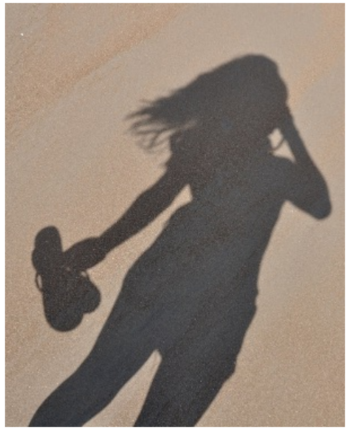


Statistical bias

μ



$\bar{x}$



How many people wash their hands after using the restroom...?

a. A study asked 6,000 randomly selected people if they wash their hands after using the restroom.

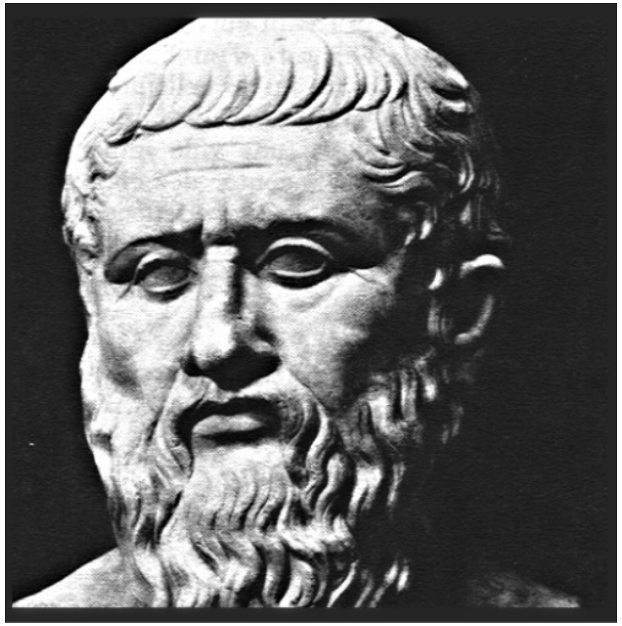
b. A study from Harris Interactive collected data by standing in public restrooms and pretending to comb their hair or put on make-up and observed whether 6,000 patrons washed their hands.

What is the parameter, and what is the statistic in these studies?

- i.e., what symbols should we use to represent the parameter and statistics in these studies?

Bias or No Bias?

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



Sad Plato says:

"Wash your hands!"

Q: How can we prevent sampling bias?

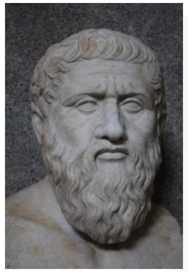
A: To prevent bias, use a **simple random sample**

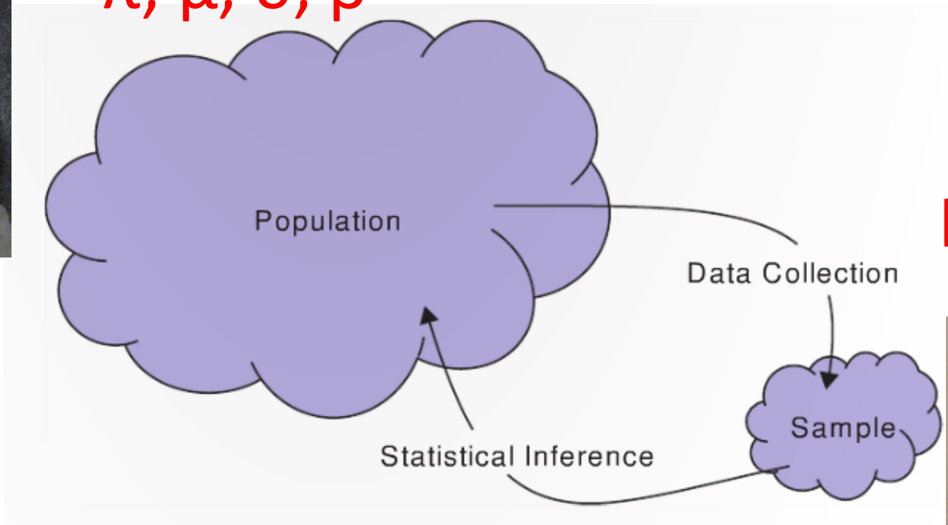
- where each member in the population is equally likely to be in the sample
  - Using a computer to do the random selection (or mechanical means)

This allows for generalizations to the population!

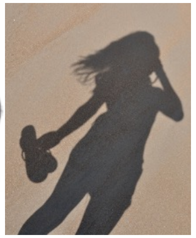
Soup analogy!

π, μ, σ, ρ





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



Avoiding bias

You need to think carefully:

What is the population I am interested in?

Does the sample reflect the population of interest?

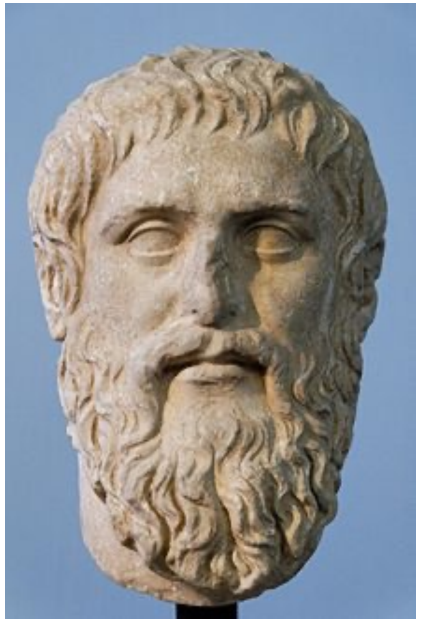
It might not be feasible to randomly select equally from all members of a population

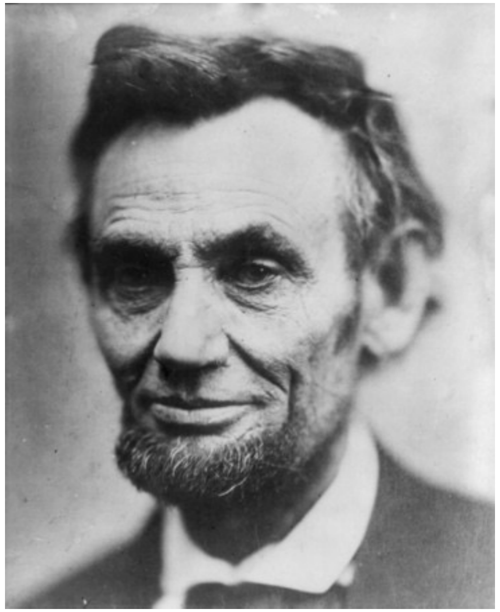
This might not be a problem as long as the sample is representative of the population.

Example: If we wanted to know proportion of people left-handed in the US, randomly sampling Yale students might be good enough.

From now on we are going to assume no bias!

Happy Plato and Lincoln



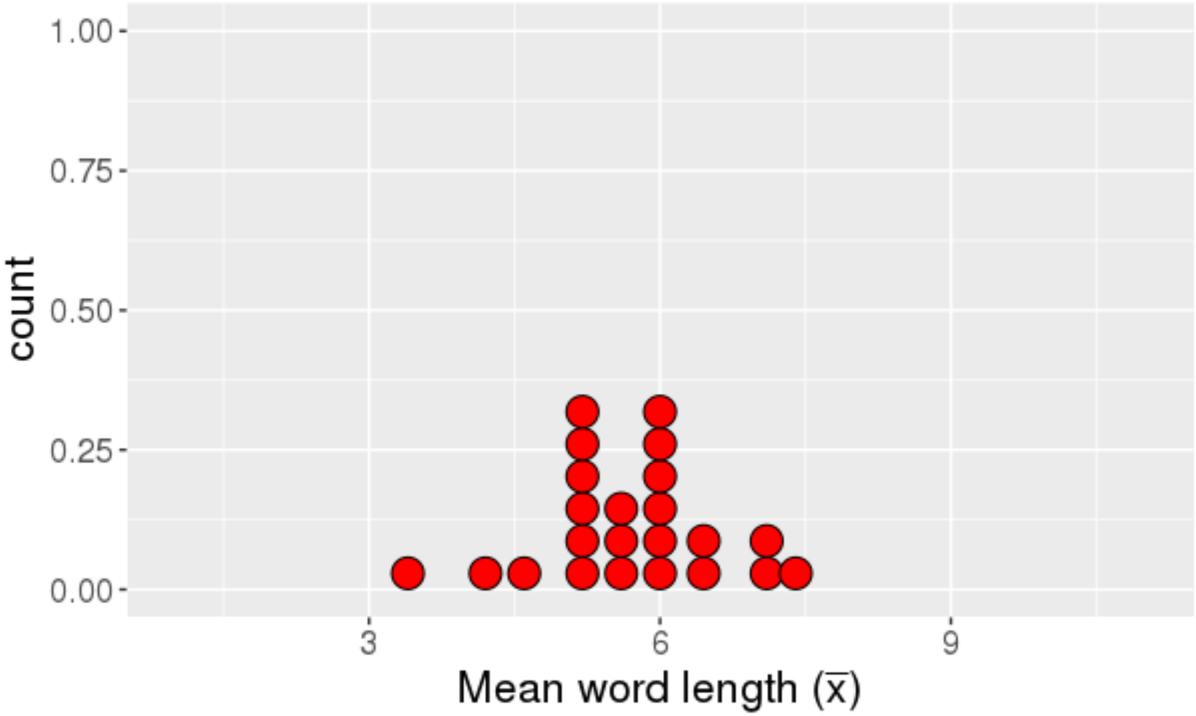


Our statistic values, on average, reflect the parameters

Sampling distributions

Recall for our distribution of Gettysburg word lengths…

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



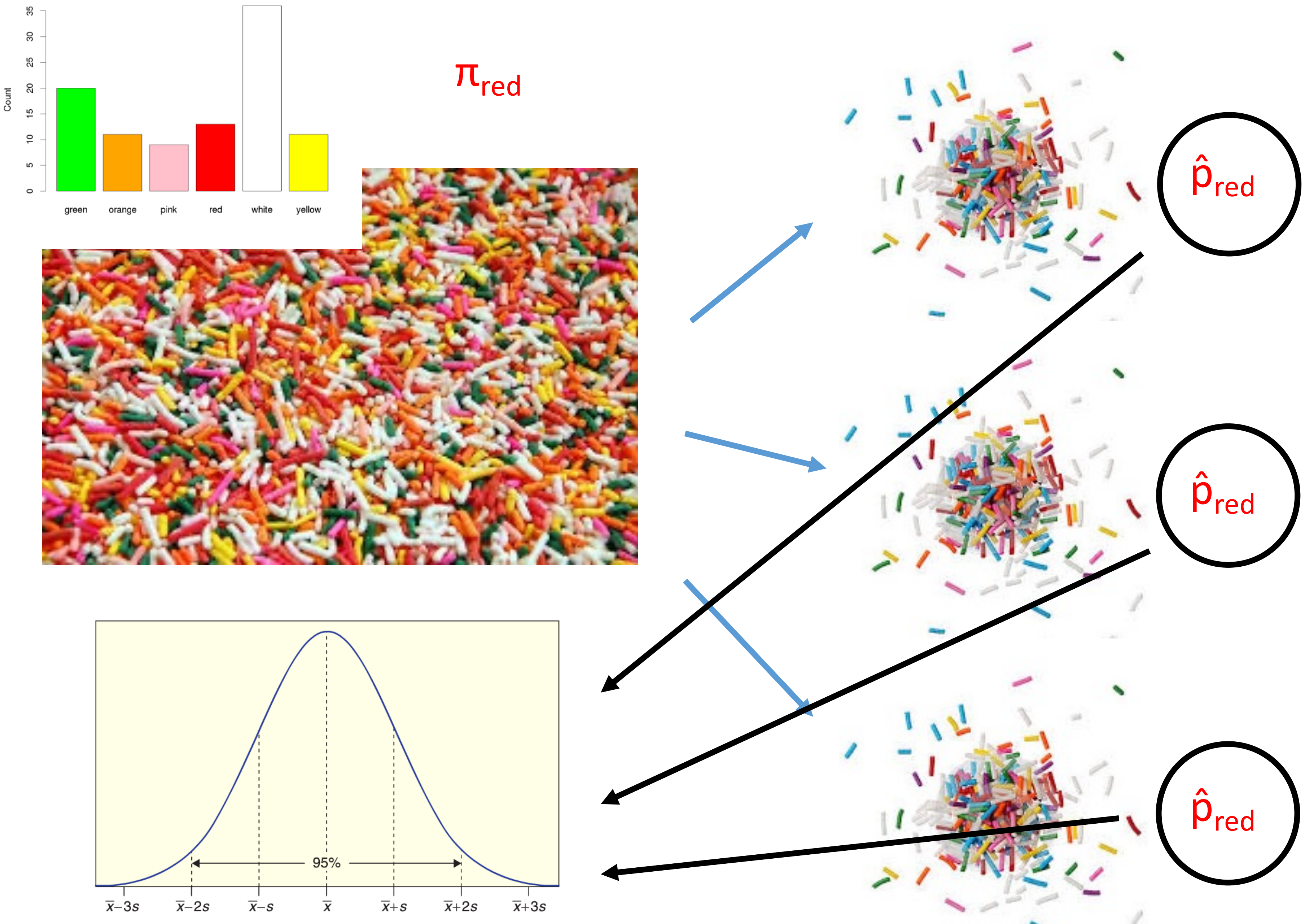
A: The mean length of 10 words ($\bar{x}$)

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

Sampling distribution

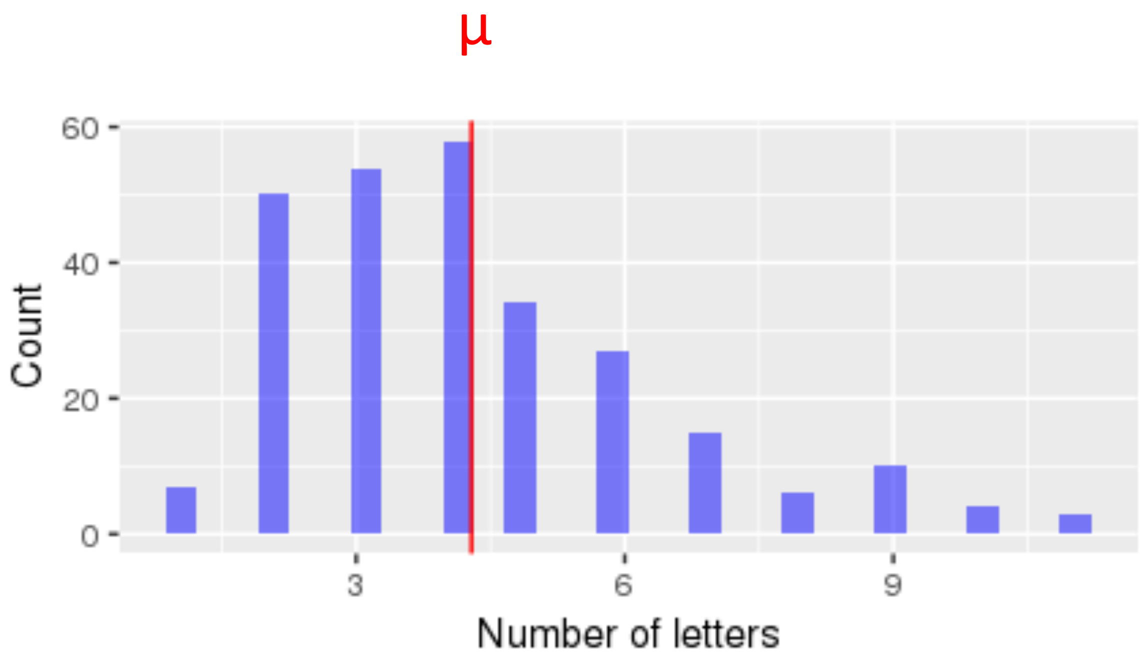
A **sampling distribution** is the distribution of <u>sample statistics</u> computed for different samples of the same size (n) from the same population.

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



Sampling distribution!

Gettysburg address word length sampling distribution



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

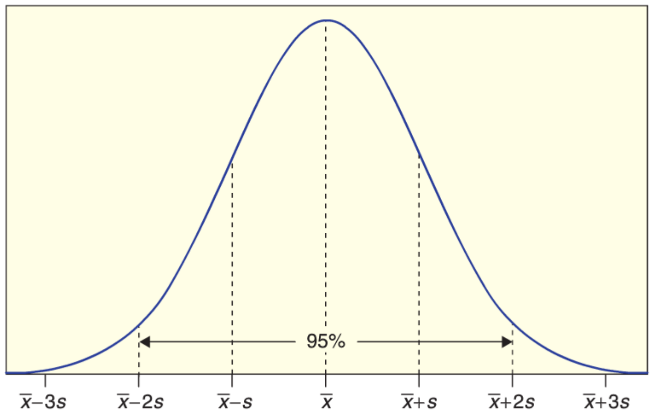
$\overline{x} = 5$

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

$\overline{x} = 4.3$

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

$\overline{x} = 4.2$



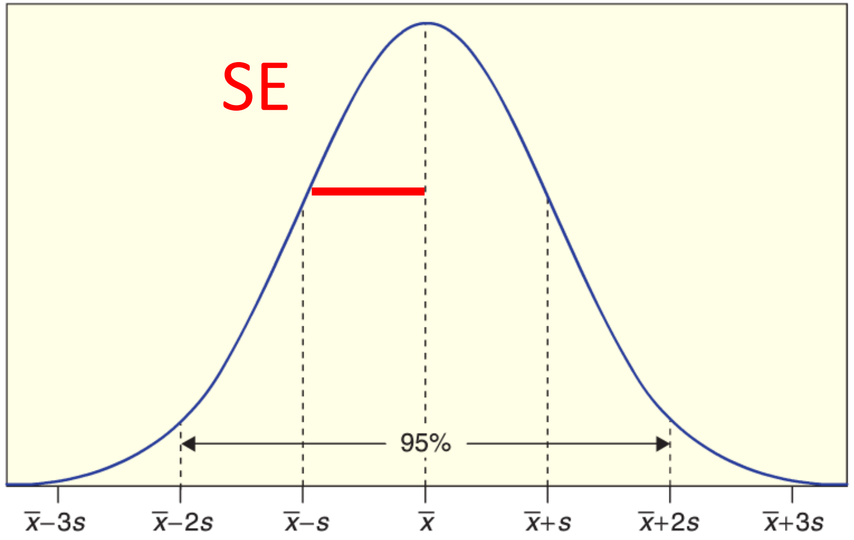
Sampling distribution!

[Gettysburg sampling distribution app]

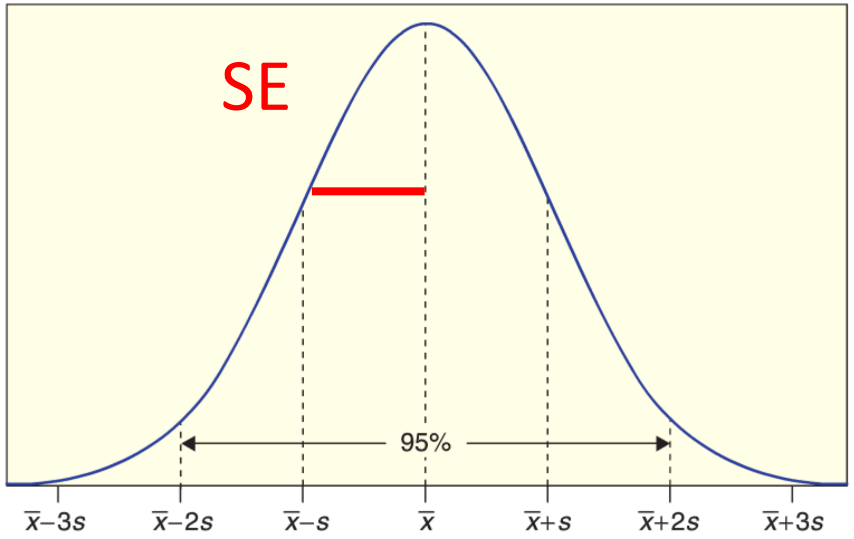
The standard error

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?



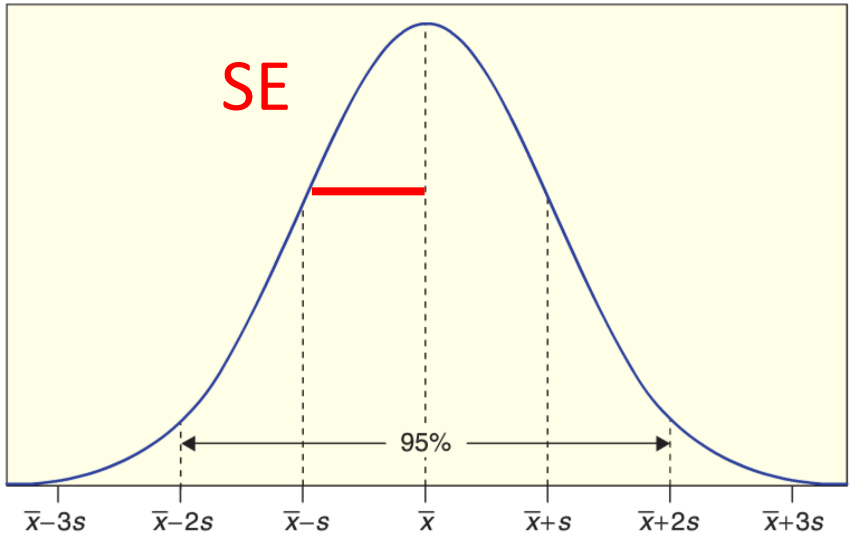
Q: If we have a large SE, would we believe a given statistic is a good estimate for the parameter?

- 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

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

Let's explore sampling distributions in R!

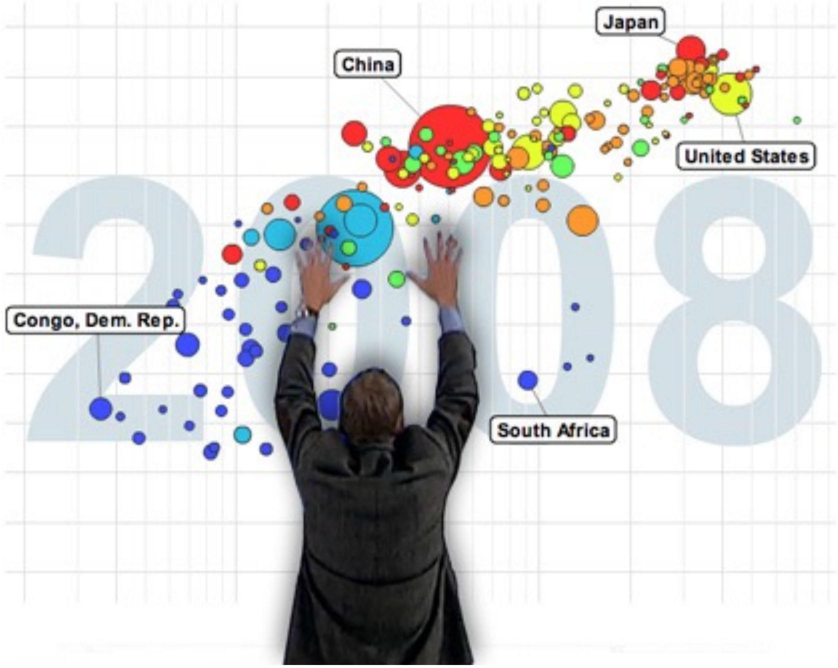
Let's create a sampling distribution in R!

Load the SDS100 library to make all SDS100 functions available

> library(SDS100)

Get the class 8 code

> download_class_code(8)



We will look at the gapminder data from 2007 which has data from countries around the world

- In particular, we will look at average life expectancy

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(lifeExp, 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

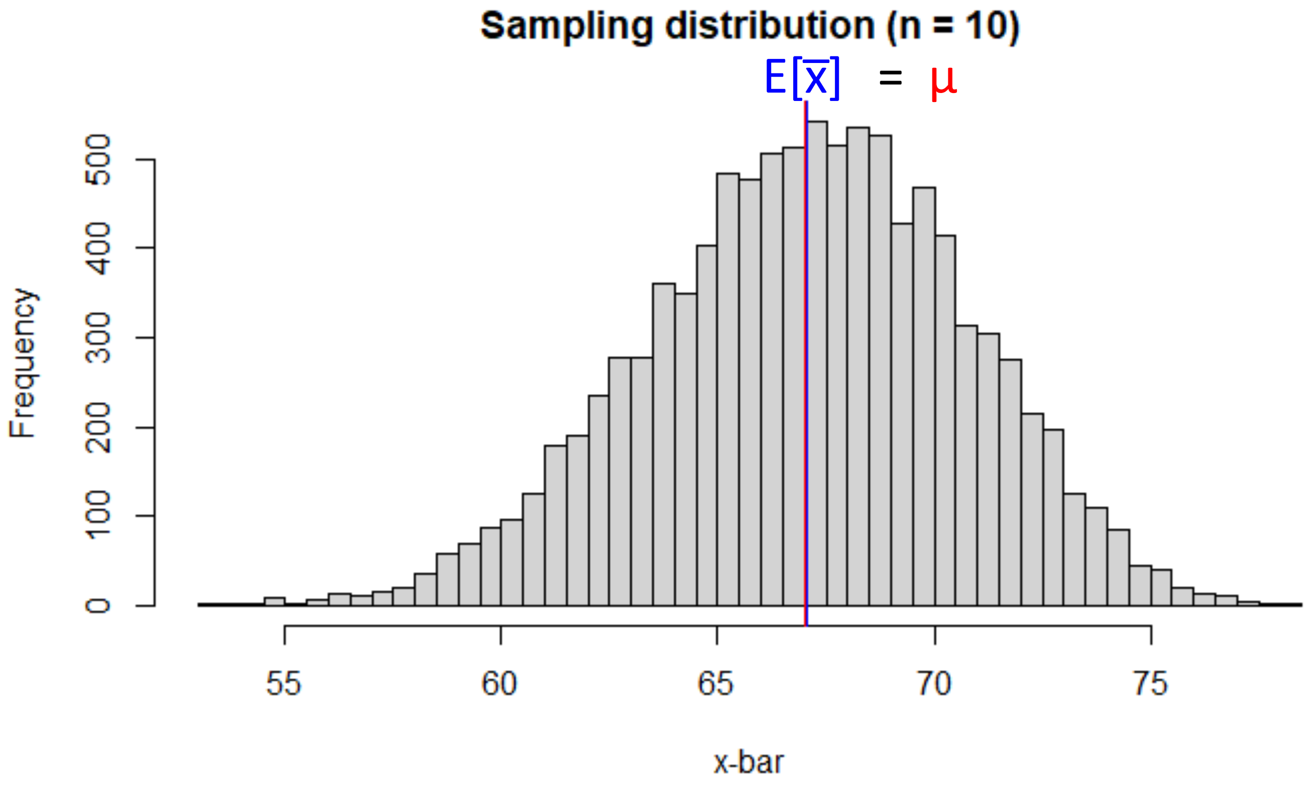
```
sampling_dist <- do_it(10000)  *  {

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

}

hist(sampling_dist)
```

Sampling distribution in R



mean(sampling_dist)
mean(lifeExp)   # 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

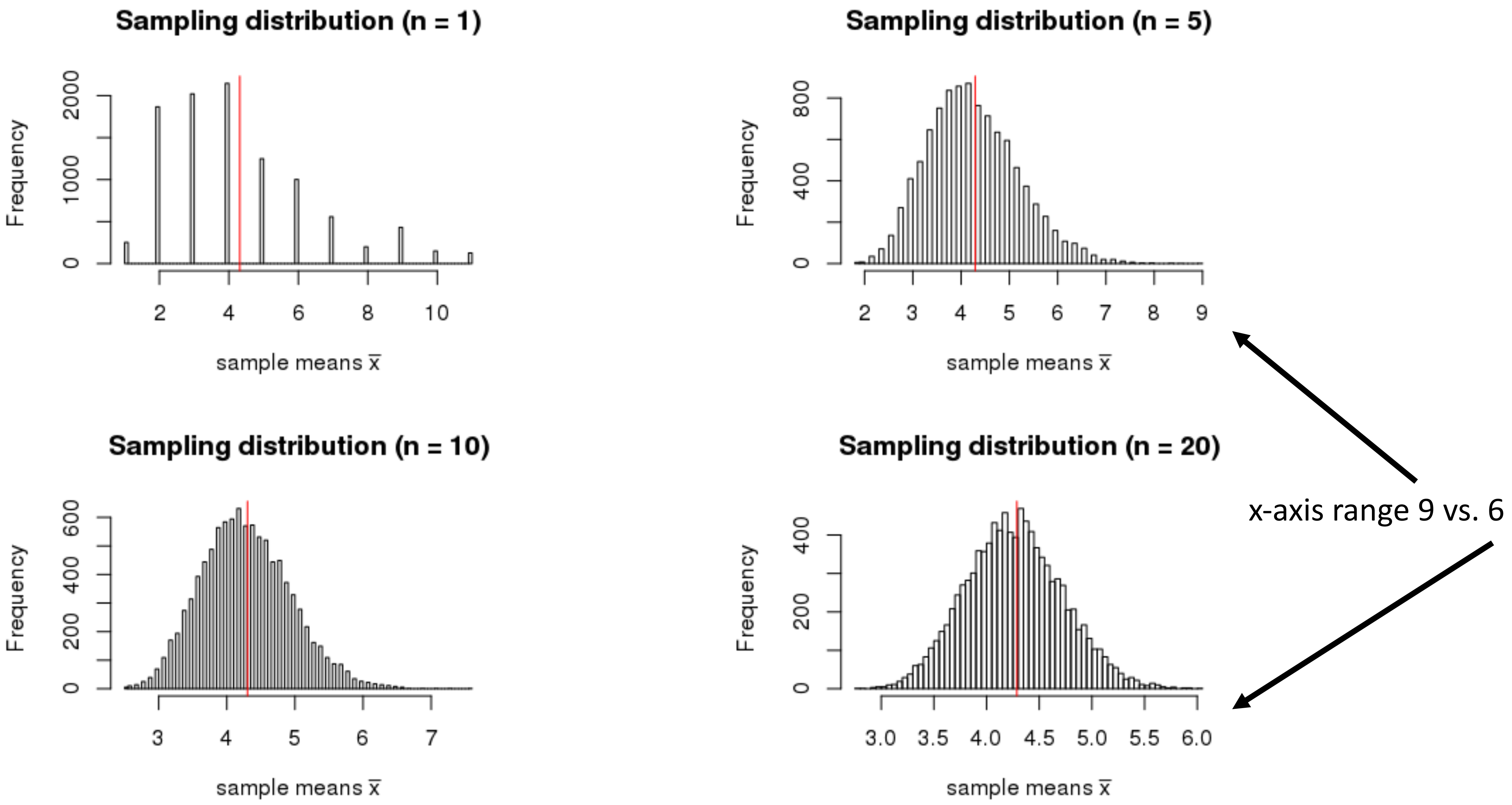
```
sampling_dist <- do_it(10000)  *  {

        curr_sample <- sample(lifeExp, 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 $E[\bar{x}] = \mu$

Note: use set.seed() for reproducibility

Sometimes it is useful to get the same sequence of random numbers

- E.g., if you want to get a consistent number when doing a random simulation

Using the set.seed(100) function can allow you to do this

- Analogous to opening a book of random numbers at page 100

Every time you call the set.seed() function you will restart the sequence of random numbers at the same place

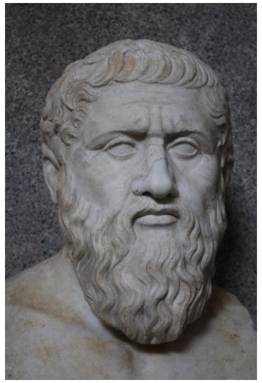
Point estimates and confidence intervals

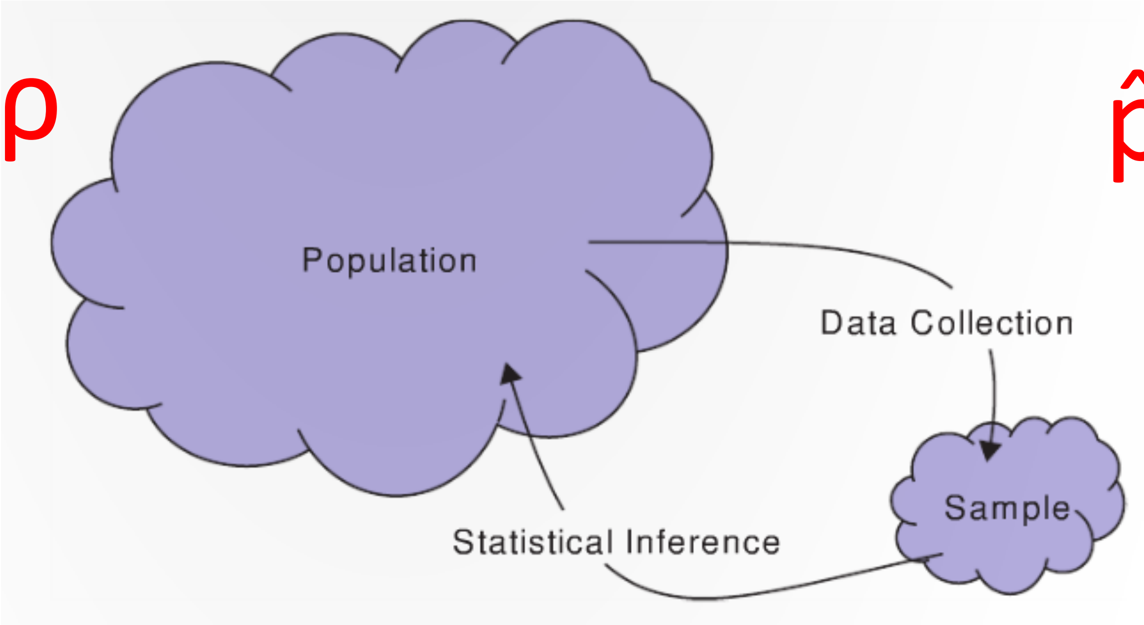
Back to the big picture: Inference

Statistical inference is…?

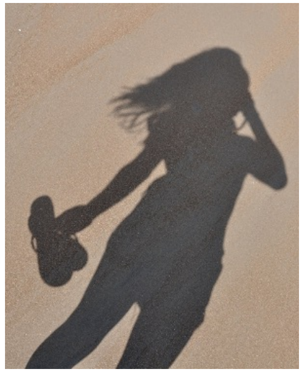
the process of drawing conclusions about the entire population based on information in a sample

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

Example: A SBU/Siena survey found that [75% of Americans] said they planned to watch the Super Bowl

Q: What are π and $\hat{p}$ here?

Q: Assuming no bias, 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 ± 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

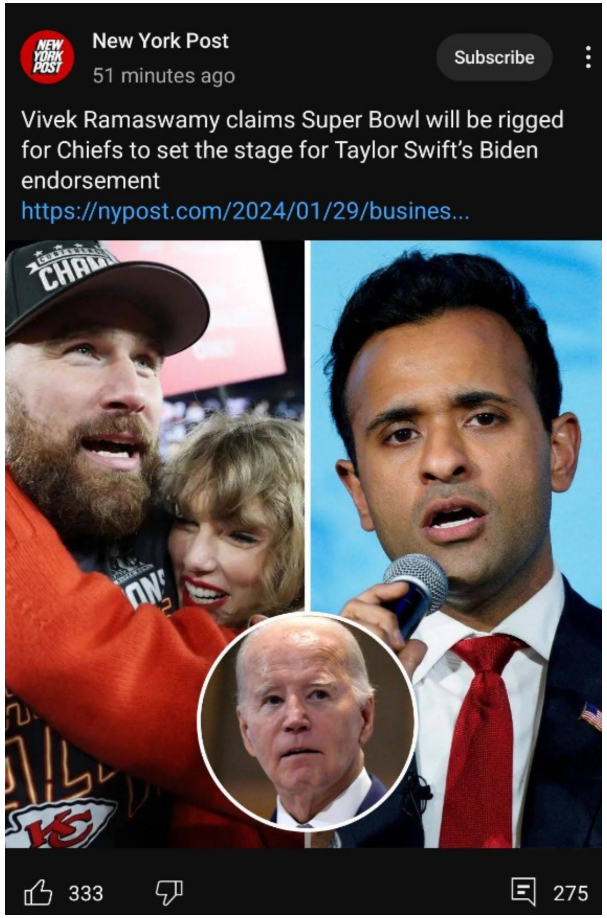
Example: Gallup poll

75% of Americans plan to watch the Super Bowl plus or minus 5%

How do we interpret this?

Says that the <u>population parameter</u> (π) lies somewhere between 70% to 80%

i.e., if they sampled all Americans, 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

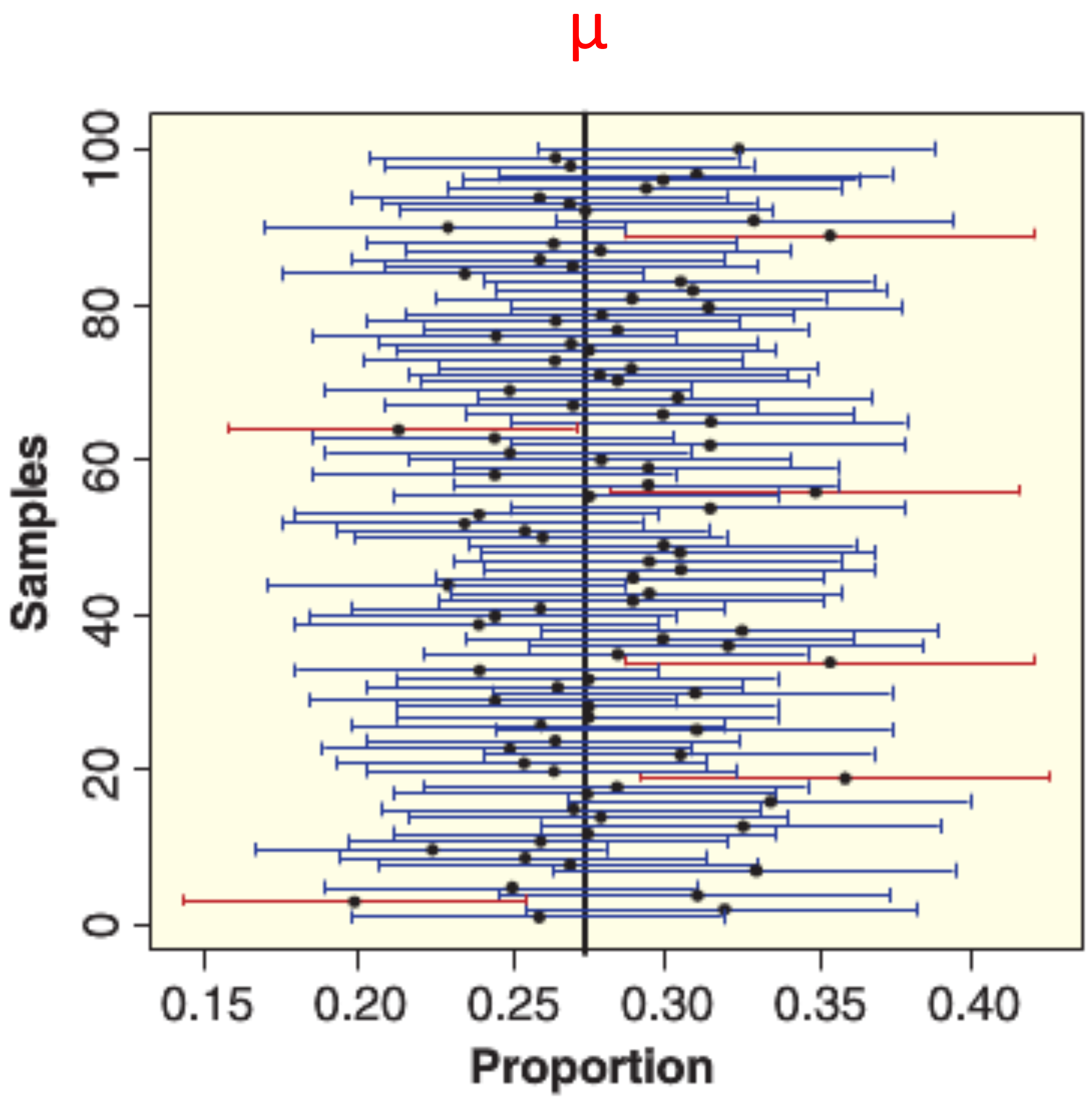
We toss intervals at it

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



For the homework (2.5c and 3.5c):
computing a 95% confidence intervals if we know the SE

To compute confidence intervals, we can use the formula:

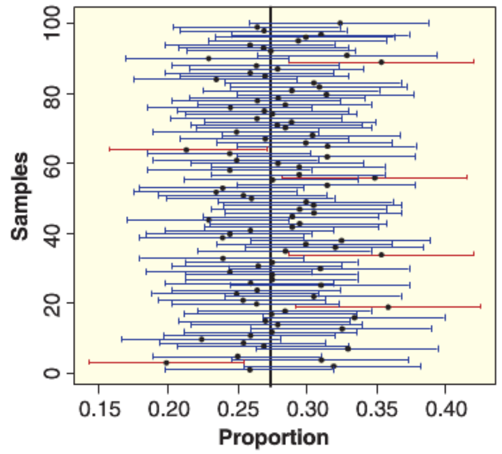
$$\text{statistic} \quad \pm \quad 2 \cdot \text{SE}$$

This is the margin of error

For example:

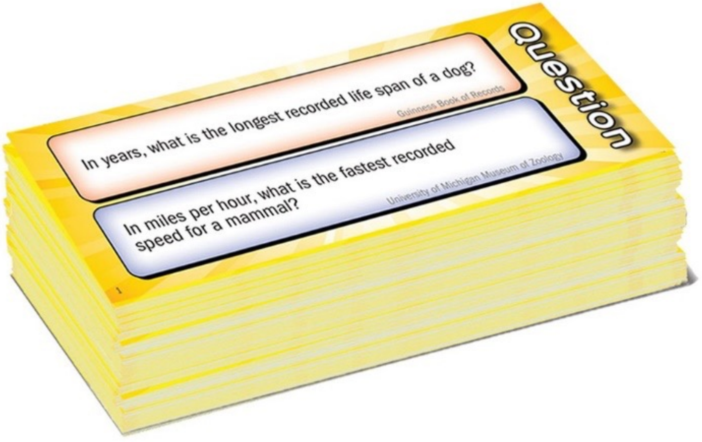
A 95% CI for μ would be: $\bar{x} \quad \pm \quad 2 \cdot \text{SE}$

A 95% CI for π would be: $\hat{p} \quad \pm \quad 2 \cdot \text{SE}$



We will explain why this formula works soon!!!

Wits and Wagers:
90% confidence interval estimator



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

Next class…

Why does this formula give a 95% CI: statistic $\pm$ $2 \cdot SE$?

How can we compute a SE from a single sample of data?