Sampling distributions, standard errors, and confidence intervals

Overview

Review: Sampling distributions and the standard error

Exploring sampling distributions in R

Confidence intervals

Announcement

Homework 4 has been posted!

It is due on Gradescope on Sunday February 15th at 11pm

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

The material this week is going to be a bit more conceptually challenging

Please attend the practice sessions and office hours to reinforce your understanding!

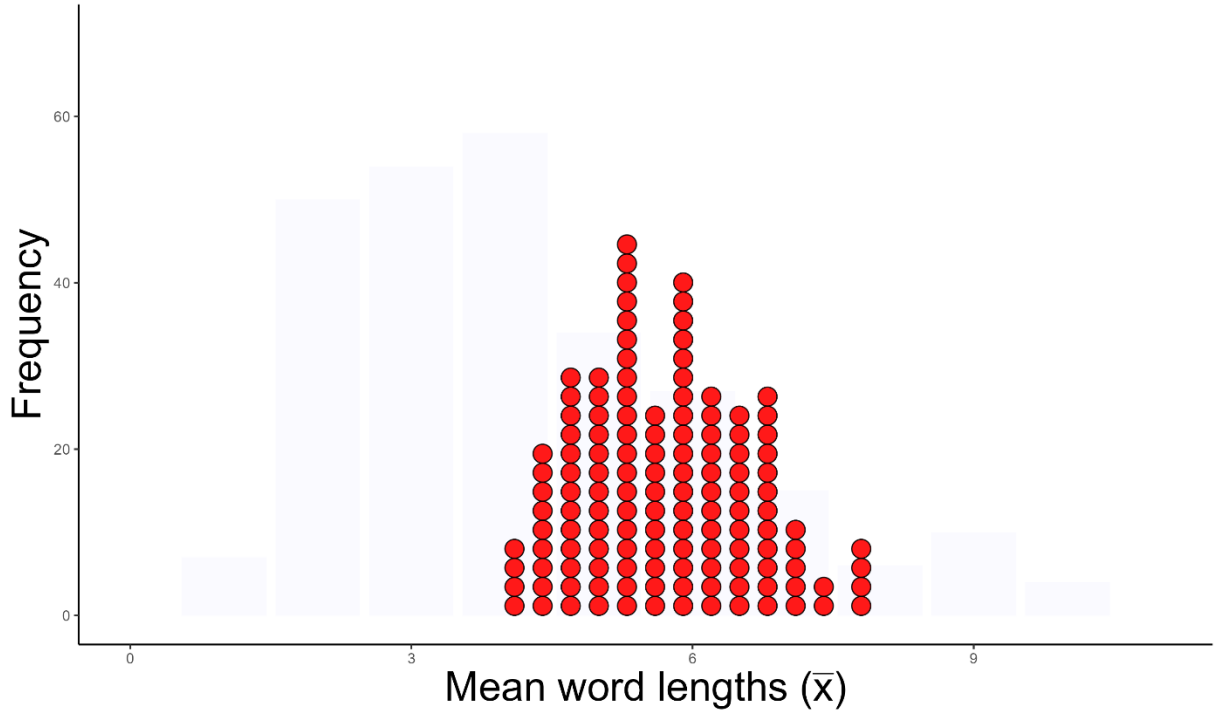
Announcement: practice session locations

Day	Time	Location
Tuesday	4, 5pm	ESC 110 (21 Sachem Street)
Wednesday	4, 5pm	1105-B Kline tower
Thursday	4, 5pm	205 Kline tower
Friday	10, 11am	1105-B Kline tower

Review: Sampling distributions

Recall for our distribution of Gettysburg word lengths…

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



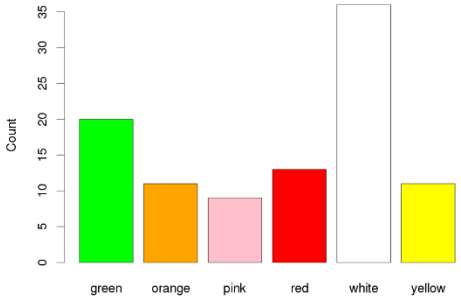
A: The mean length of 10 words ($\bar{x}$)

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

Sampling distribution

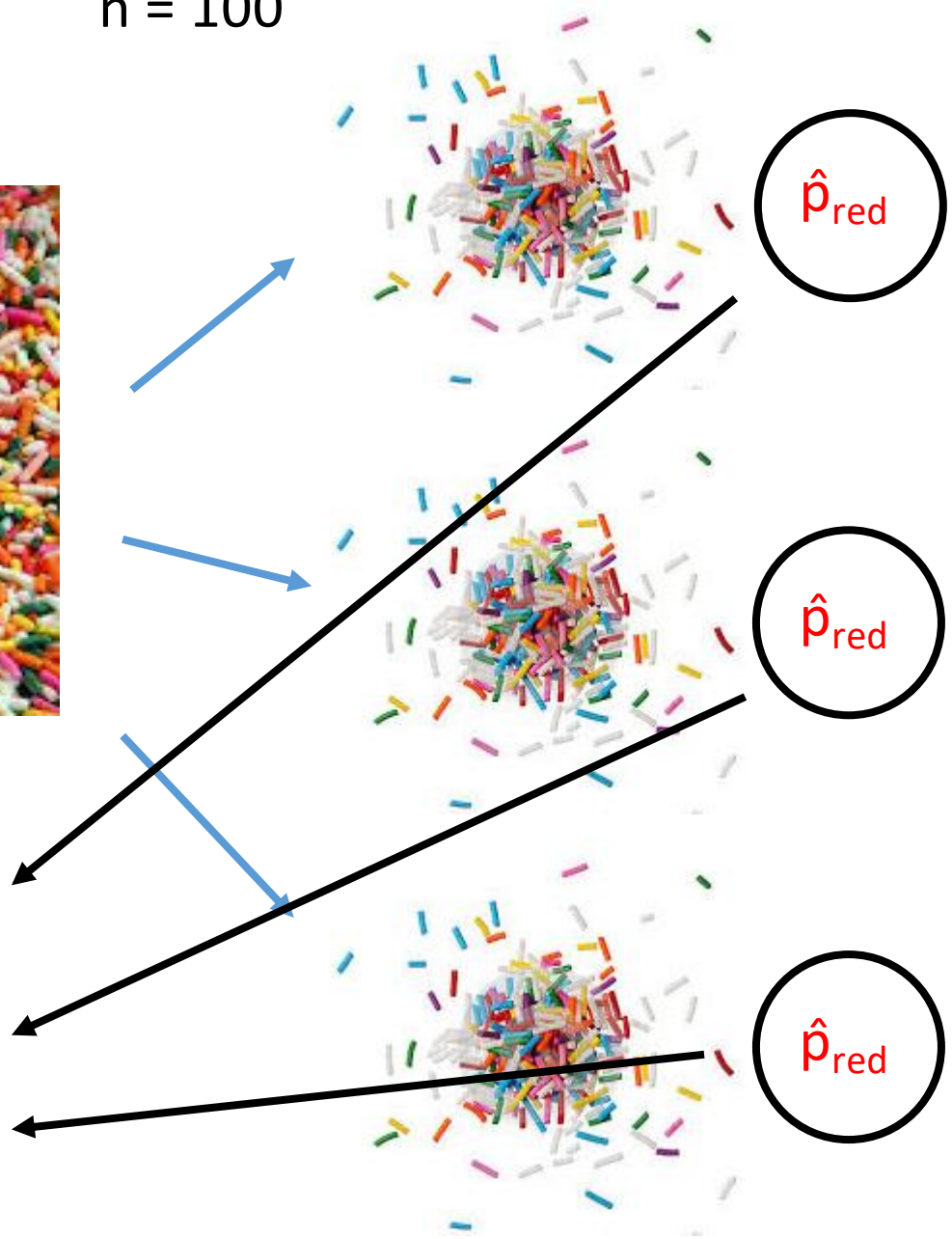
A **sampling distribution** is the distribution of sample statistics computed from different samples of the same size (n) from the same population

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

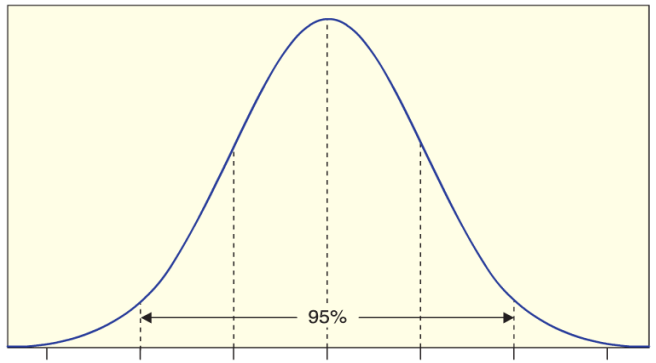


π_{red}

n = 100



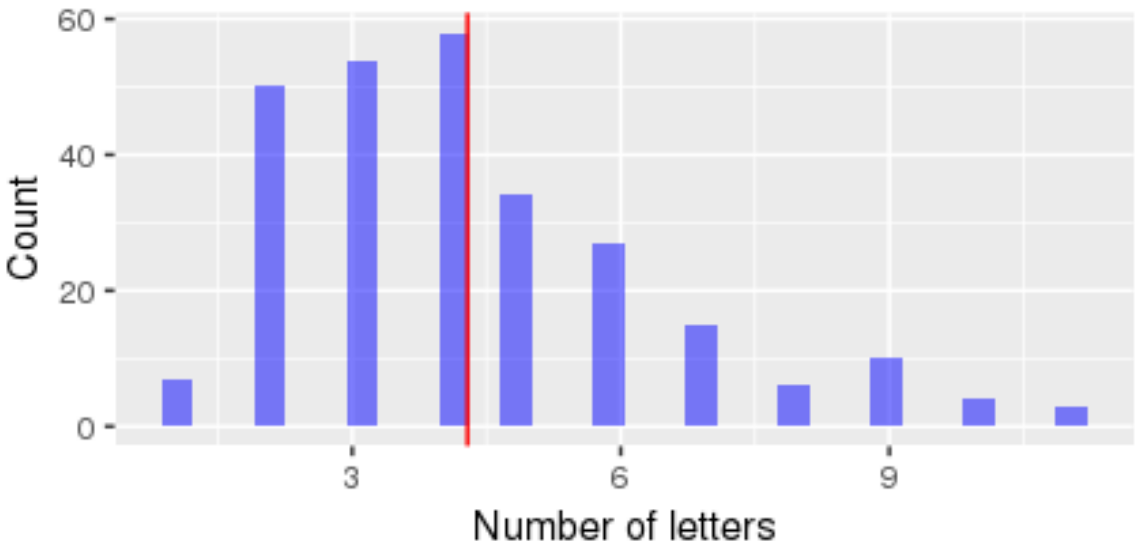
Distribution of $\hat{p}$'s



Sampling distribution!

Gettysburg address word length sampling distribution

μ



n = 10

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

$\bar{x} = 5$

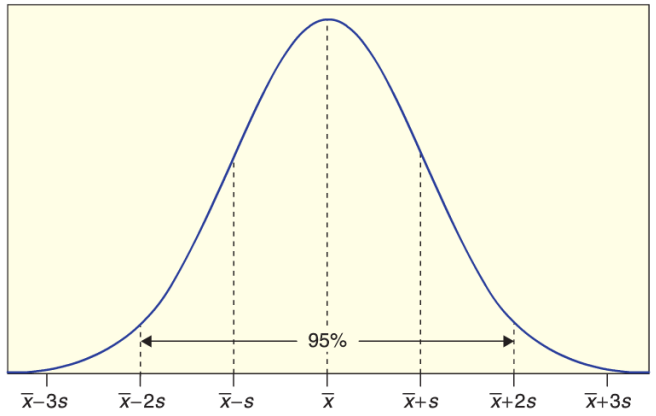
2, 6, 2, 6, 6,
2, 5, 3, 2, 9

$\bar{x} = 4.3$

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

$\bar{x} = 4.2$

Distribution of $\bar{x}$'s



Sampling distribution!

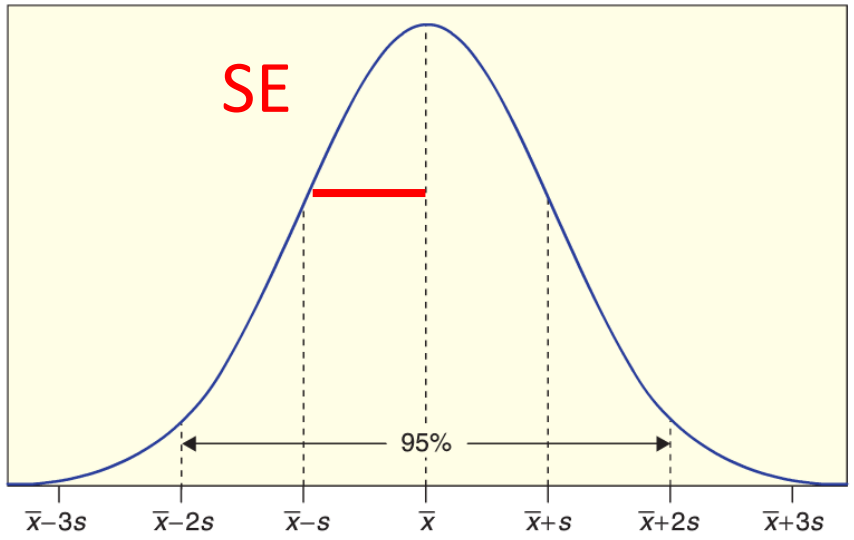
Gettysburg sampling distribution app

The standard error

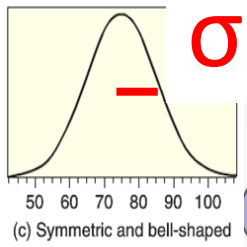
The **standard error** of a statistic, denoted SE, is the standard deviation of the sample statistic

- i.e., SE is the standard deviation of the *sampling distribution*

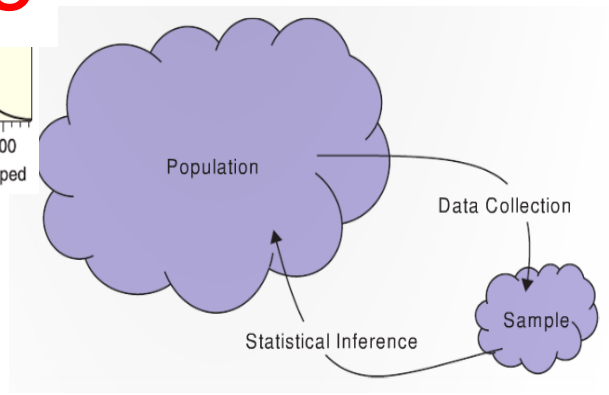
Note: **standard error** SE is different from σ and s!

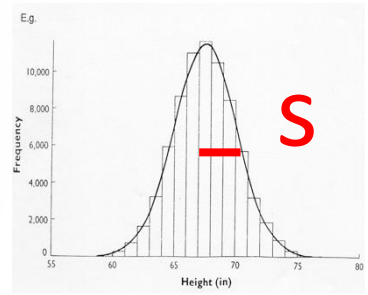


Sampling distribution!



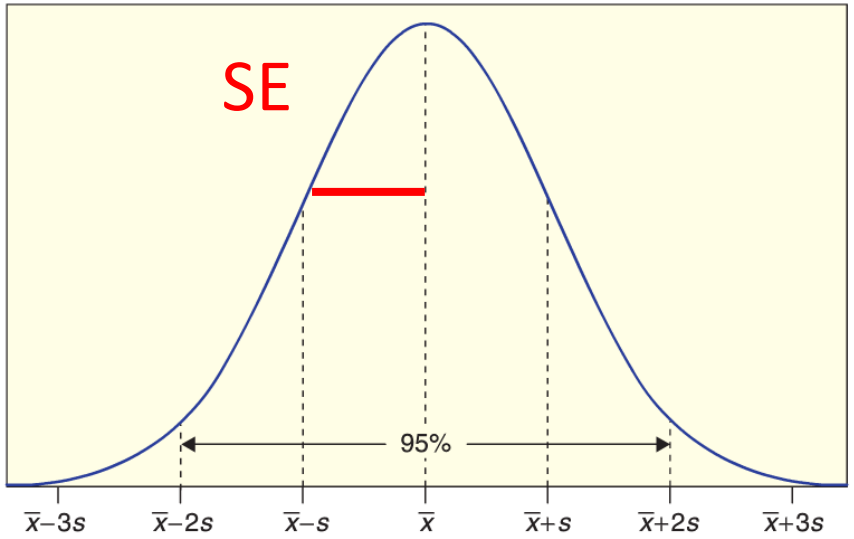
σ





s

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 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

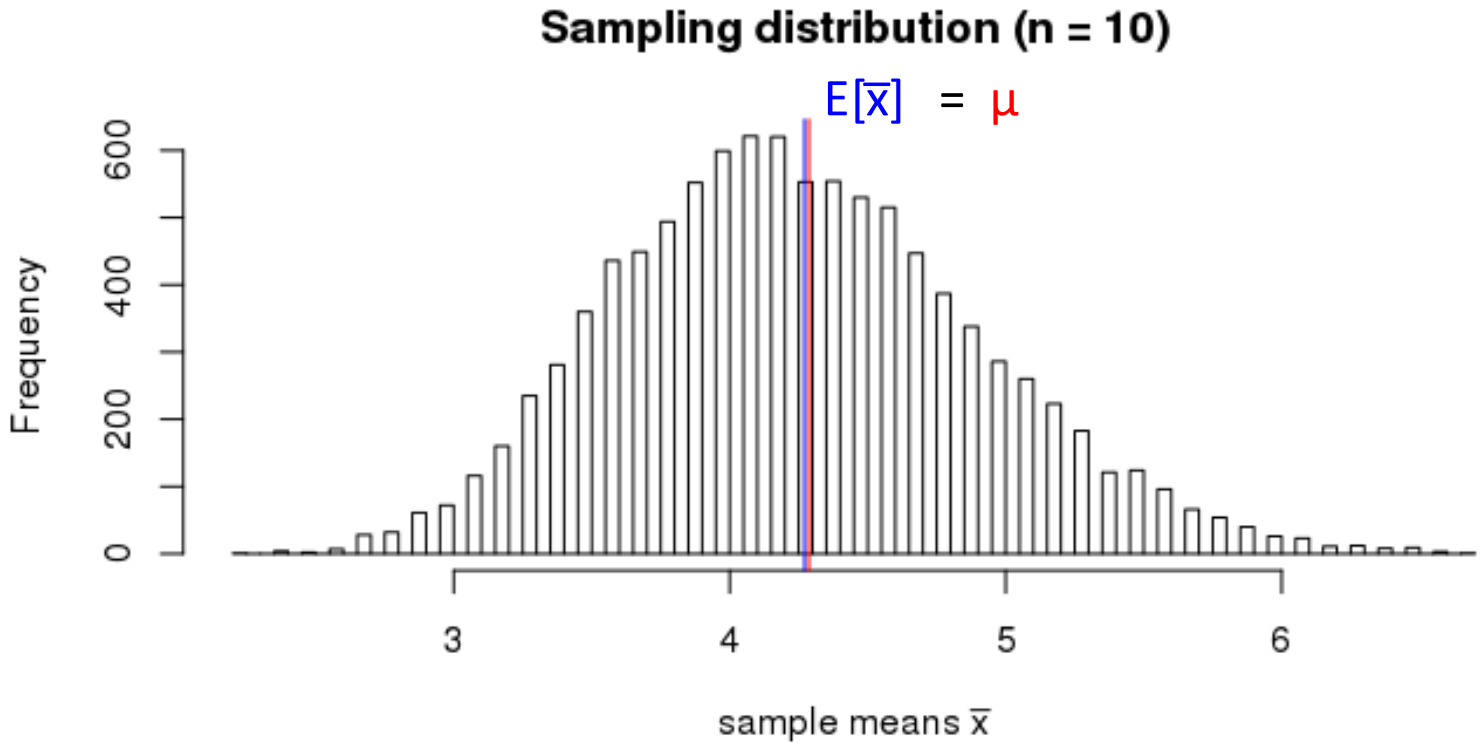
- It returns a vector with the values created each time the code is repeated

```
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)
SE <- sd(sampling_dist)
```

Sampling distribution in R




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

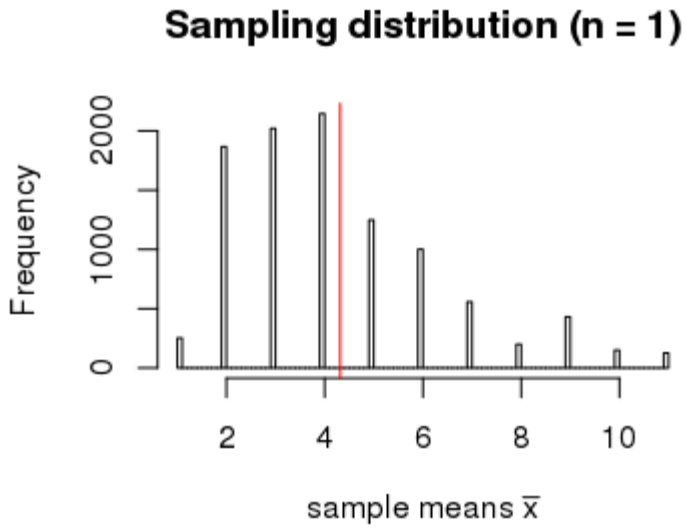
Changing the sample size n

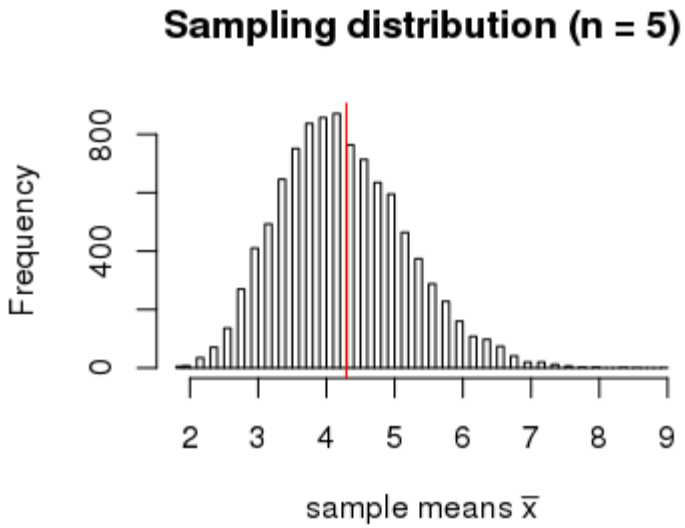
What happens to the sampling distribution as we change ***n***?

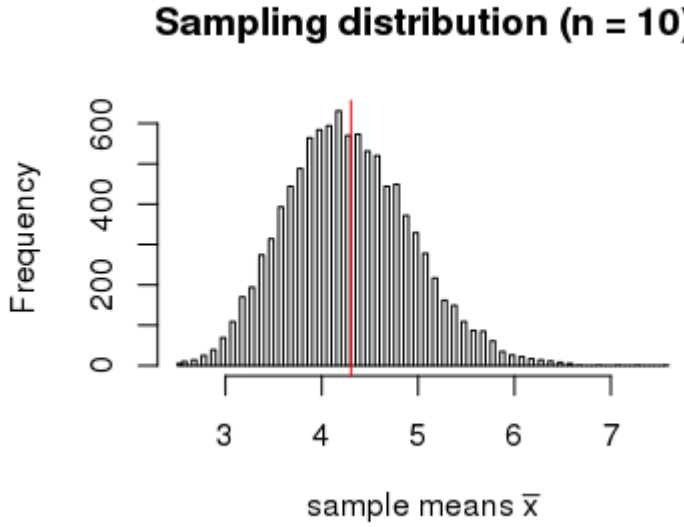
- Experiment for n = 1, 5, 20, 80

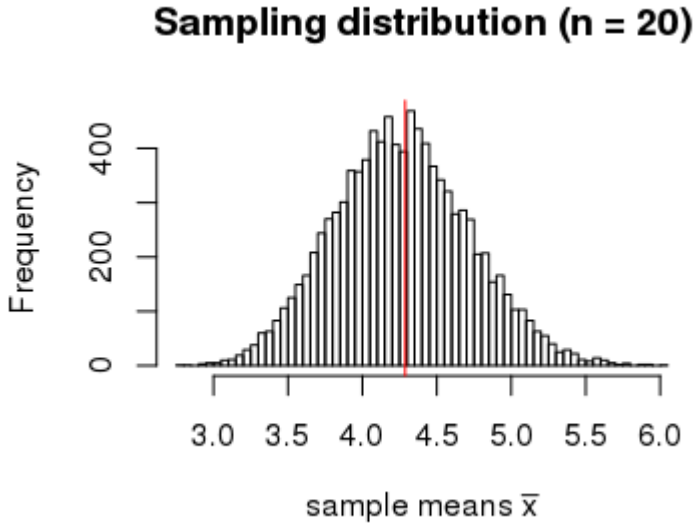
```
sampling_dist <- do_it(10000)  *  {

        curr_sample <- sample(word_lengths, 20)

        mean(curr_sample)

}

hist(sample_means, breaks = 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 $E[\bar{x}] = \mu$

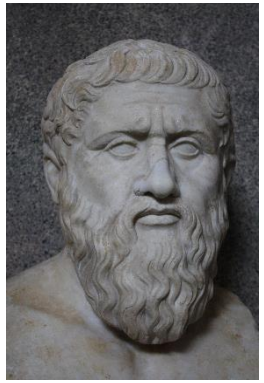
Point estimates and confidence intervals

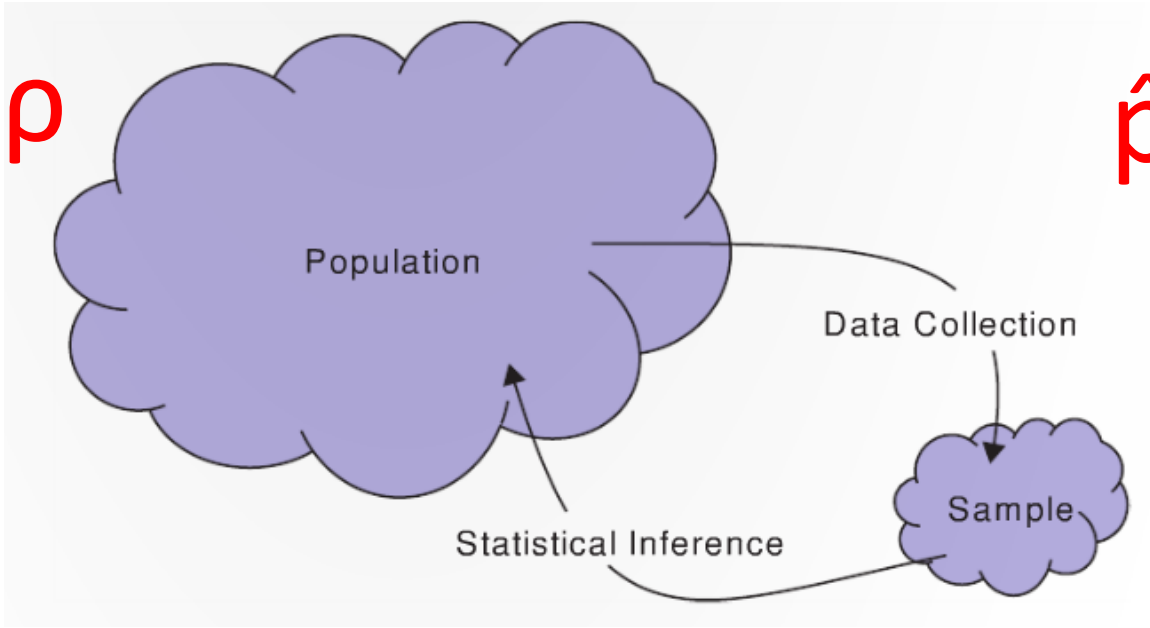
Back to the big picture: Inference

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 Americans approve of Trump's job performance according to a poll of 2,464 people

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

(outdated graphic)

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 likely be in this range

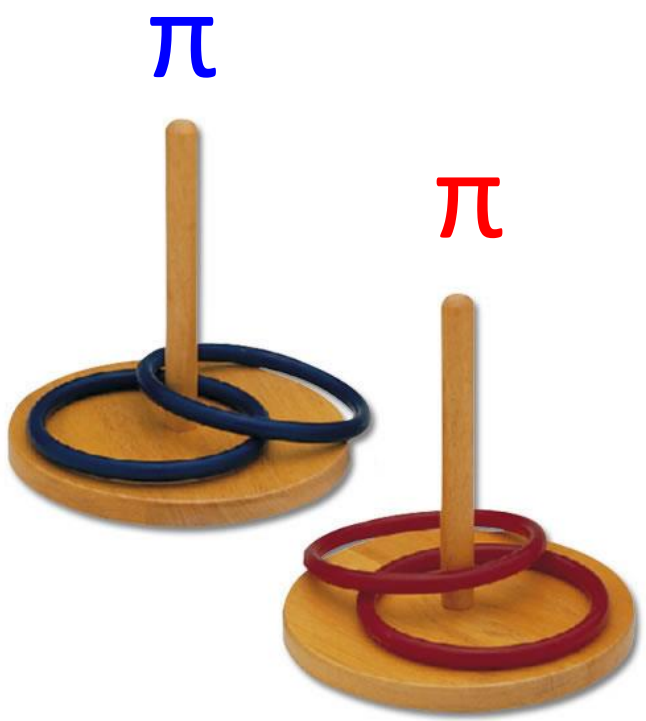
Confidence Intervals

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…



A parameter exists in the world

We toss intervals at it

- These are our confidence intervals

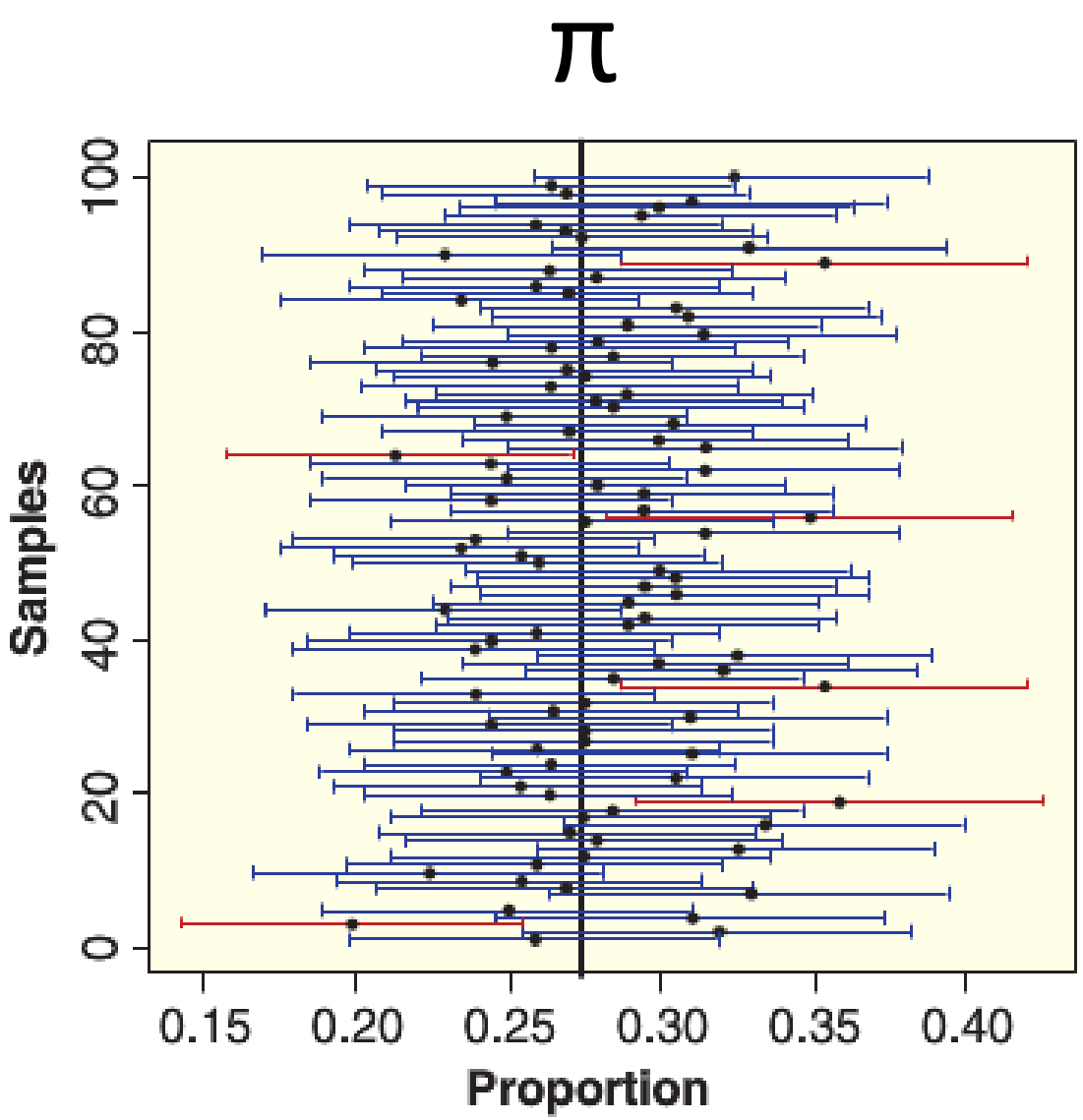
95% of those intervals capture the parameter

- (95% is our confidence level)

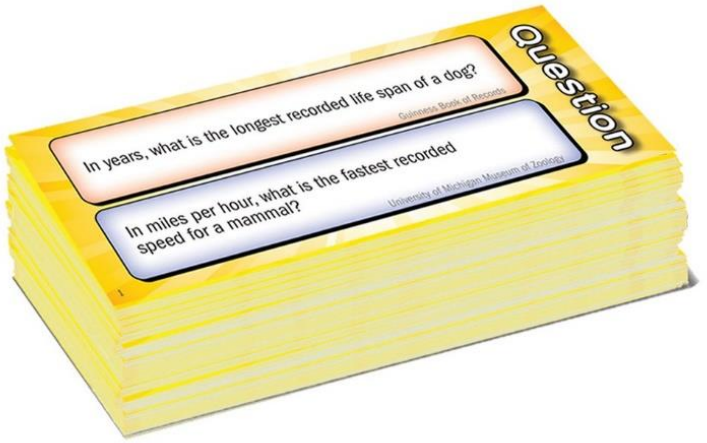
Confidence Intervals

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

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



Wits and Wagers:
90% confidence interval estimator



I will ask 10 questions that have numeric answers

Please come up with a range of values that contains the true value 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]