Elementary Statistics – Principles of Confidence Intervals

Dr. Ab Mosca (they/them)

Slides based off slides courtesy of Kaitlyn Cook (https://www.smith.edu/people/kaitlyn-cook)

Plan for Today

- Confidence Intervals
  - Random variables
  - Sampling distributions

Warm Up: Probability

Sample space is the collection (set) of *all possible outcomes* of a random experiment. (Denoted with S)

An ***event space*** is a collection of possible outcomes

General Addition Rule: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Practice: We flip a fair coin three times. Let A be the event that the coin lands on heads on the first two flips or the last 2 flips. What is $P(A)$?

Hint: What is S? What is A? What is B? What is $A \cap B$?

Random Variables

Notation can become unwieldy as events we're interested in get more complicated.

Ex. Imagine trying to write the event that we "flip a coin 100 times and get exactly 50 heads" using intersections and unions

$A_1 =$ "$50\ heads\ on\ the\ first\ 50\ rolls$"
$A_2 =$ "$50\ heads\ on\ the\ 2nd\ - 51st\ rolls$"
...

It would take forever to break all this down

Random Variables

Notation can become unwieldy as events we're interested in get more complicated.

Ex. Imagine trying to write the event that we "flip a coin 100 times and get exactly 50 heads" using intersections and unions

$A_1 =$ "$50\ heads\ on\ the\ first\ 50\ rolls$"
$A_2 =$ "$50\ heads\ on\ the\ 2nd\ - 51st\ rolls$"
...

It would take forever to break all this down

Instead, we use a special numerical representation to make this easier – ***Random Variables***

Random Variables

Random Variables provide a numerical representation for our events of interest, which makes systematically describing their probabilities easier.

Notation:

We use capital letters to denote random variables and lowercase letters to denote their realizations:

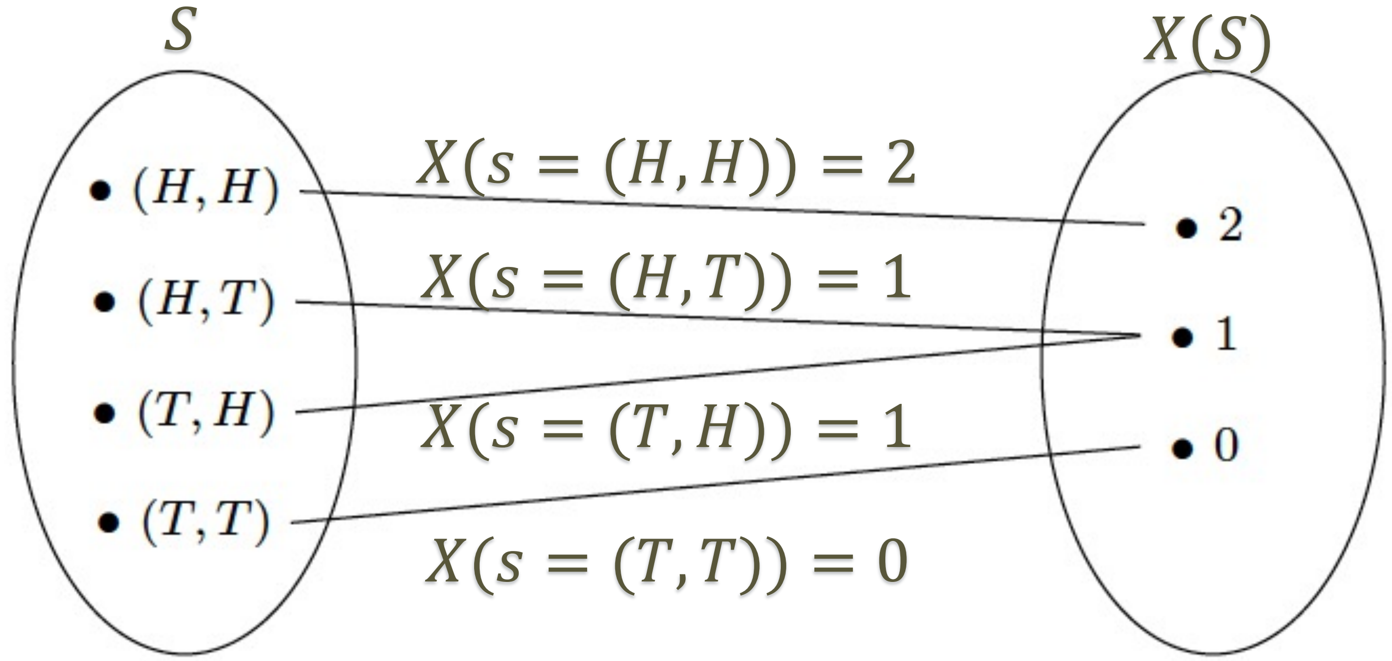
- Let s be an outcome in our sample space, S
- The random variable, $X(s)$ is a function that maps outcomes in our sample space to real numbers

Random Variables

Let s be an outcome in our sample space, S
The ***random variable, $X(s)$*** is a function that maps outcomes in our sample space to real numbers

Ex. Suppose we flip a coin twice (this is our event, s).
We can define the random variable,

$$X(s) = the\ number\ of\ heads\ in\ s$$

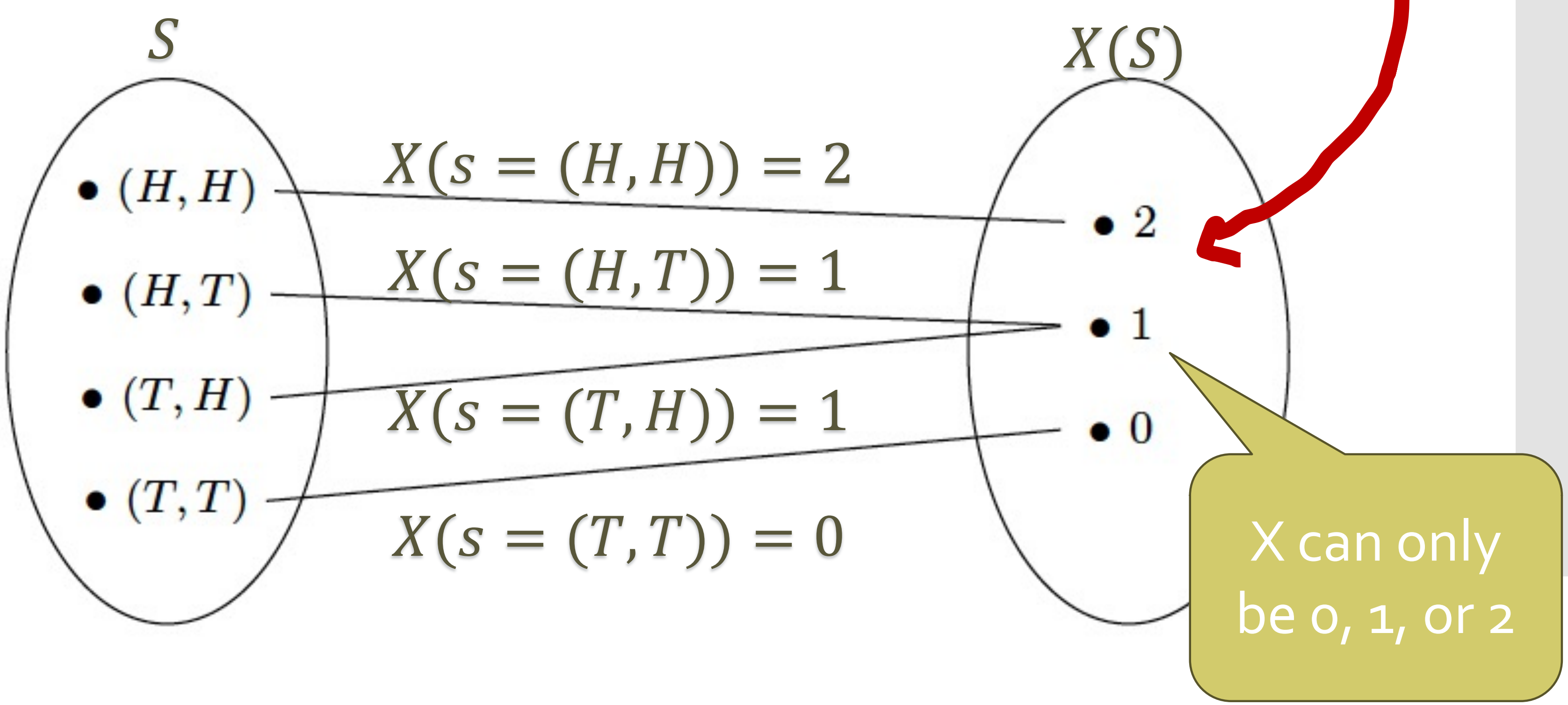


Discrete Random Variables

A ***discrete random variable, $X(s)$*** can only take on particular values in an interval.

Ex. Suppose we flip a coin twice (this is our event, s).
We can define the random variable,

$$X(s) = the\ number\ of\ heads\ in\ s$$



Discrete Random Variables

A ***discrete random variable,*** $\boldsymbol{X(s)}$ can only take on particular values in an interval.

The ***probability distribution*** for a discrete random variable, X lists:

1. All of the values, x, that X might take on
2. The probability that X takes on each of those values

Discrete Random Variables

A ***discrete random variable,*** $\boldsymbol{X(s)}$ can only take on particular values in an interval.

The ***probability distribution*** for a discrete random variable, X lists:

1. All of the values, x, that X might take on
2. The probability that X takes on each of those values

Ex. Suppose we flip a coin twice (this is our event, s).

$$X(s) = the\ number\ of\ heads\ in\ s$$

Outcome	(H, H)	(H, T)	(T, H)	(T, T)
$X(s)$	2	1	1	0
Probability	0.25	0.25	0.25	0.25

x	0	1	2
$P(X = x)$	0.25	0.50	0.25

A ***discrete random variable,*** $\boldsymbol{X(s)}$ can only take on particular values in an interval.

The ***probability distribution*** for a discrete random variable, X lists:

1. All of the values, x, that X might take on
2. The probability that X takes on each of those values

Ex. Suppose we flip a coin three times (this is our event, s).

$$X(s) = the\ number\ of\ heads\ in\ s$$

Outcome	?	…	…	?
$X(s)$				
Probability	?	?	?	?

Practice:

(1) What are the possible outcomes for s?

(2) What is $X(s)$ for each outcome?

(3) What is the probability of each outcome?

Practice:

(1) What are the possible values for x?

(2) What is the probability of each value?

A ***discrete random variable***, $\boldsymbol{X(s)}$ can only take on particular values in an interval.

The ***probability distribution*** for a discrete random variable, X lists:

1. All of the values, x, that X might take on
2. The probability that X takes on each of those values

Ex. Suppose we flip a coin three times (this is our event, s).

$$X(s) = the\ number\ of\ heads\ in\ s$$

Outcome	?	…	…	?
$X(s)$				
Probability	?	?	?	?

↓

x	?	…	?
$P(X = x)$	?	?	?

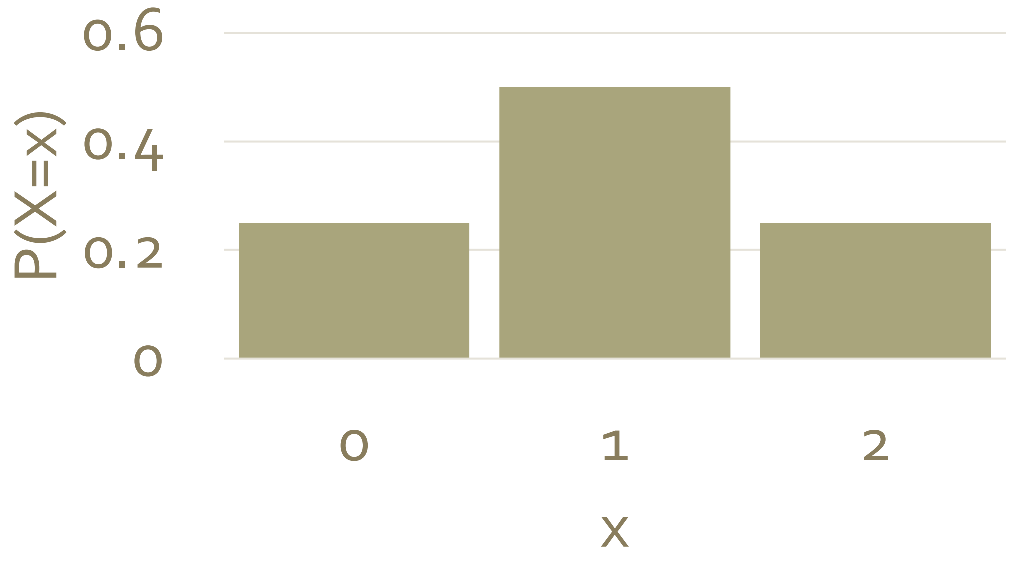
Probability Mass Functions

A ***probability mass function (PFM)*** is a way to visualize the the probability distribution for a discrete random variable.

Ex. Suppose we flip a coin twice (this is our event, s).

$$X(s) = the\ number\ of\ heads\ in\ s$$

x	0	1	2
$P(X = x)$	0.25	0.50	0.25



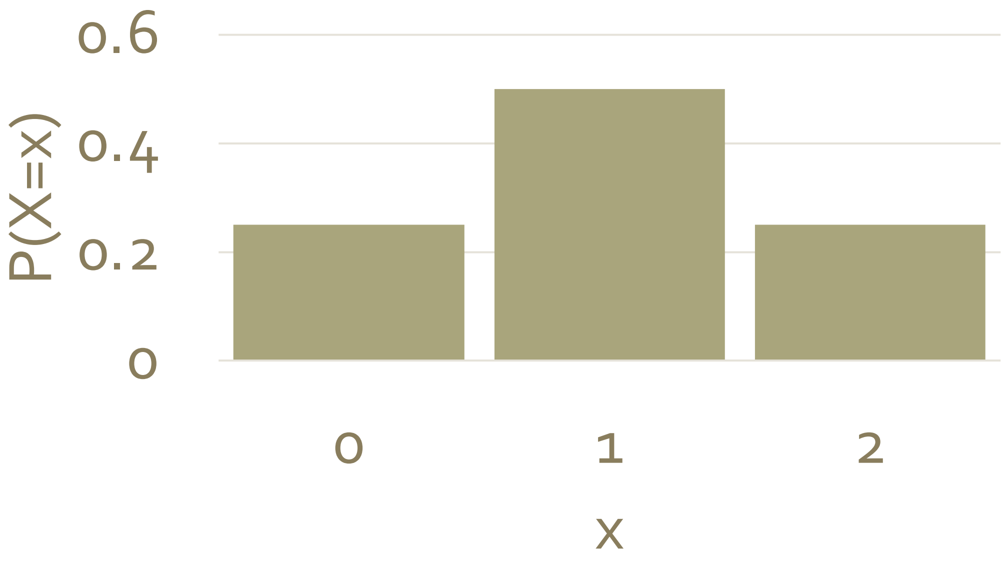
Probability Mass

A ***probability mass function (PFM)*** is a way to visualize the the probability distribution for a discrete random variable.

Ex. Suppose we flip a coin twice (this is our event, s).

$$X(s) = the\ number\ of\ heads\ in\ s$$

x	0	1	2
$P(X = x)$	0.25	0.50	0.25



Practice:

Suppose we flip a coin three times (this is our event, s).

$$X(s) = the\ number\ of\ heads\ in\ s$$

Visualize the probability mass function for X

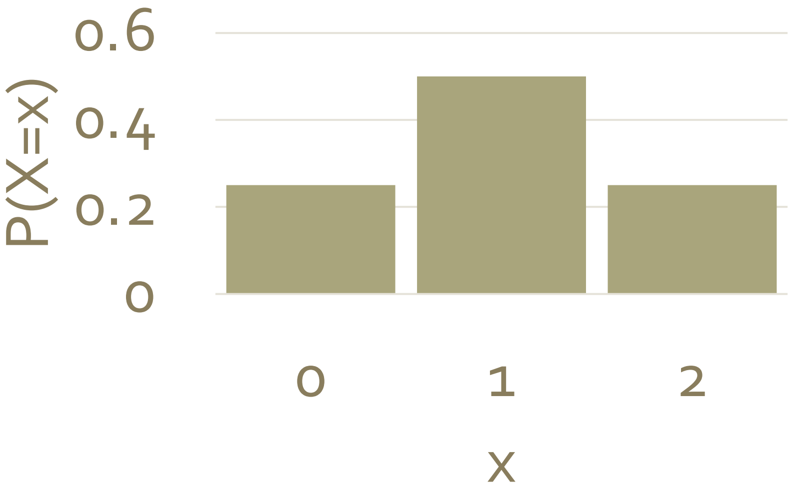
Probability Mass Functions

A ***probability mass function (PFM)*** is a way to visualize the the probability distribution for a discrete random variable.

Ex. Suppose we flip a coin twice (this is our event, s).

$$X(s) = the\ number\ of\ heads\ in\ s$$

x	0	1	2
$P(X = x)$	0.25	0.50	0.25



$$P(X = x) = \begin{cases} 0.25, x = 0 \\ 0.5, x = 1 \\ 0.25, x = 2 \end{cases}$$

Bernoulli Distributions

If our random variable only takes on two values, one indicating "success" ($x = 1$) and the other indicating "failure" ($x = 0$), then our random variable has a ***Bernoulli distribution***.

Notation:

- $X \sim Bern(p)$, where p is a population parameter representing the probability of success

Practice:
What is p (probability of success) for each of the following?

Ex. Let $X(s) = whether\ a\ coin\ toss\ lands\ on\ heads$

Ex. Let $X(s) = whether\ a\ die\ lands\ on\ 6$

Bernoulli Distributions

Given a random variable with a Bernoulli distribution, we can model the number of successes ($X = 1$) across n independent trials (called Bernoulli trials) using the ***Binomial distribution***.

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

p is a population parameter representing the probability of success

$\binom{n}{x}$ is called a Binomial coefficient, and is read "n choose x"

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}, n! = n\times(n-1)\times(n-2)\times\cdots\times 2\times 1$$

Random Variables and Statistics

Why do we care about random variables?

Recall: We want to know about populations, but we only have sample data to work with.
So we estimate population parameters using ***sample statistics***.

- Sample mean: $\bar{x}$
- Sample proportion: $\hat{p}$

Ex.

➔ In an experiment we sample 500 individuals and ask them about their weight gain during pregnancy
  ➔ The mean of all their answers is our sample mean, $\bar{x}$, which is our estimate for the population

➔ In an experiment we sample 3,010 movies and assess whether or not they made over $100 million at the box office
  ➔ The mean of all their answers is our sample proportion, $\hat{p}$, which is our estimate for the population

Random Variables and Statistics

Why do we care about random variables?

Recall: We want to know about populations, but we only have sample data to work with.
So we estimate population parameters using ***sample statistics***.

- Sample mean: $\bar{x}$
- Sample proportion: $\hat{p}$

Turns out, our **sample statistics are random variables**!

Random Variables and Statistics

Why do we care about random variables?

Recall: We want to know about populations, but we only have sample data to work with.
So we estimate population parameters using ***sample statistics***.

- Sample mean: $\bar{x}$
- Sample proportion: $\hat{p}$

Turns out, our **sample statistics are random variables**!

Ex. Suppose we want to estimate the probability, p, that a six-sided die lands on a six when rolled once.

Let's each do an experiment. Grab a die, roll it 20 times and record the number of times you roll a 6. Record your proportion of 6's on the board.

Random Variables and Statistics

Ex. Suppose we want to estimate the probability, p, that a six-sided die lands on a six when rolled once.

Let's each do an experiment. Grab a die, roll it 20 times and record the number of times you roll a 6. Record your proportion of 6's on the board.

Across repeated experiments (*of the same sample size from the same population*) our sample estimate will vary.

Looking at the distribution of sample estimates allows us to quantify how much we expect our sample estimate to vary.

Random Variables and Statistics

Ex. Suppose we want to estimate the probability, p, that a six-sided die lands on a six when rolled once.

Let's each do an experiment. Grab a die, roll it 20 times and record the number of times you roll a 6. Record your proportion of 6's on the board.

Across repeated experiments (*of the same sample size from the same population*) our sample estimate will vary.

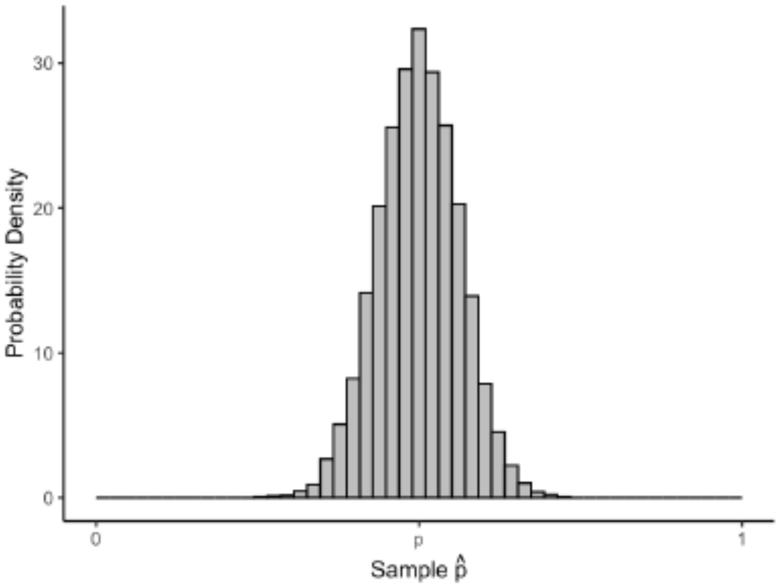
Looking at the distribution of sample estimates allows us to quantify how much we expect our sample estimate to vary.

In reality, we cannot take hundreds of samples to construct a probability distribution. Instead, a method called ***Bootstrapping*** is used. It results in a probability distribution that approximates the actual distribution very closely.

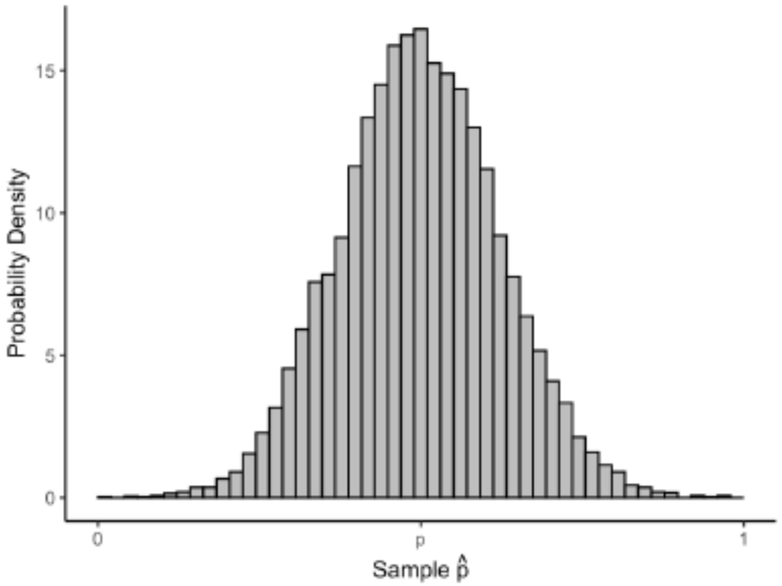
Sampling Distributions

A ***sampling distribution*** is the distribution of all possible values of a sample statistic from *samples of a given size* (n) from a *given population*.

- Tells us how our sample statistic varies from one sample to another
- Mean of the distribution is the true population parameter, p
- Spread gives us a sense of what range of values to expect for $\hat{p}$
  - Depends on the distribution of X_i and n



$\hat{p}$ likely to be close to p

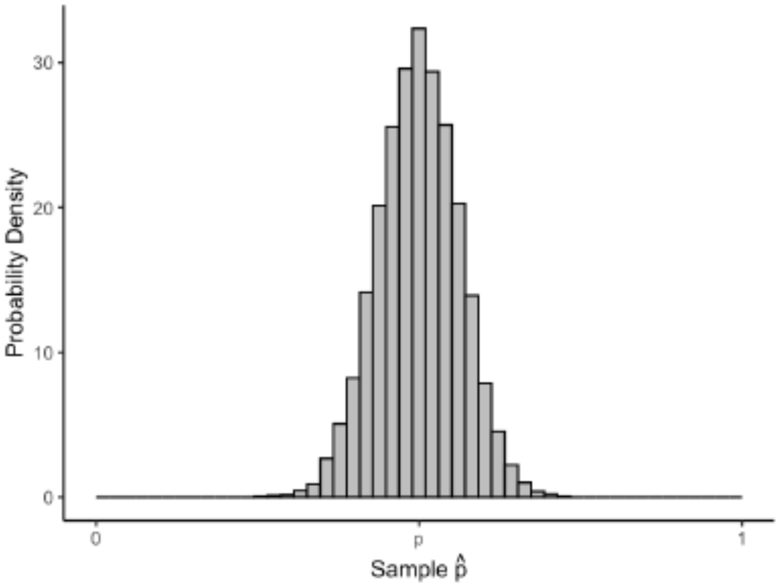


less certain that any given $\hat{p}$ is close to p

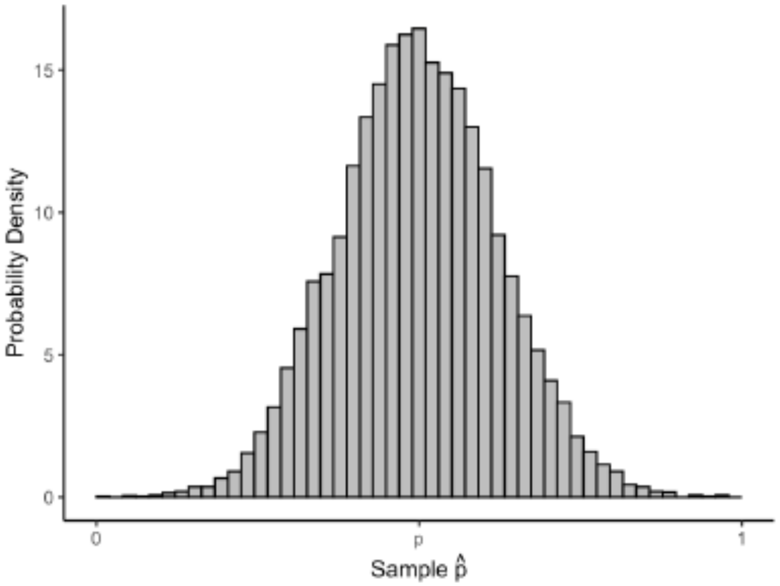
Sampling Distributions

We approximate the ***sampling distribution*** using the ***bootstrap*** method.

- Approximates how our sample statistic varies from one sample to another
- Spread gives us a (good) sense of what range of values to expect for $\hat{p}$



$\hat{p}$ likely to be close to p



less certain that any given $\hat{p}$ is close to p

Recapping

What we know:

- Our sample estimates are random variables
- Given samples of the same size (n) from the same population, they will vary
- The sampling distribution visualizes this variance for us

Recapping

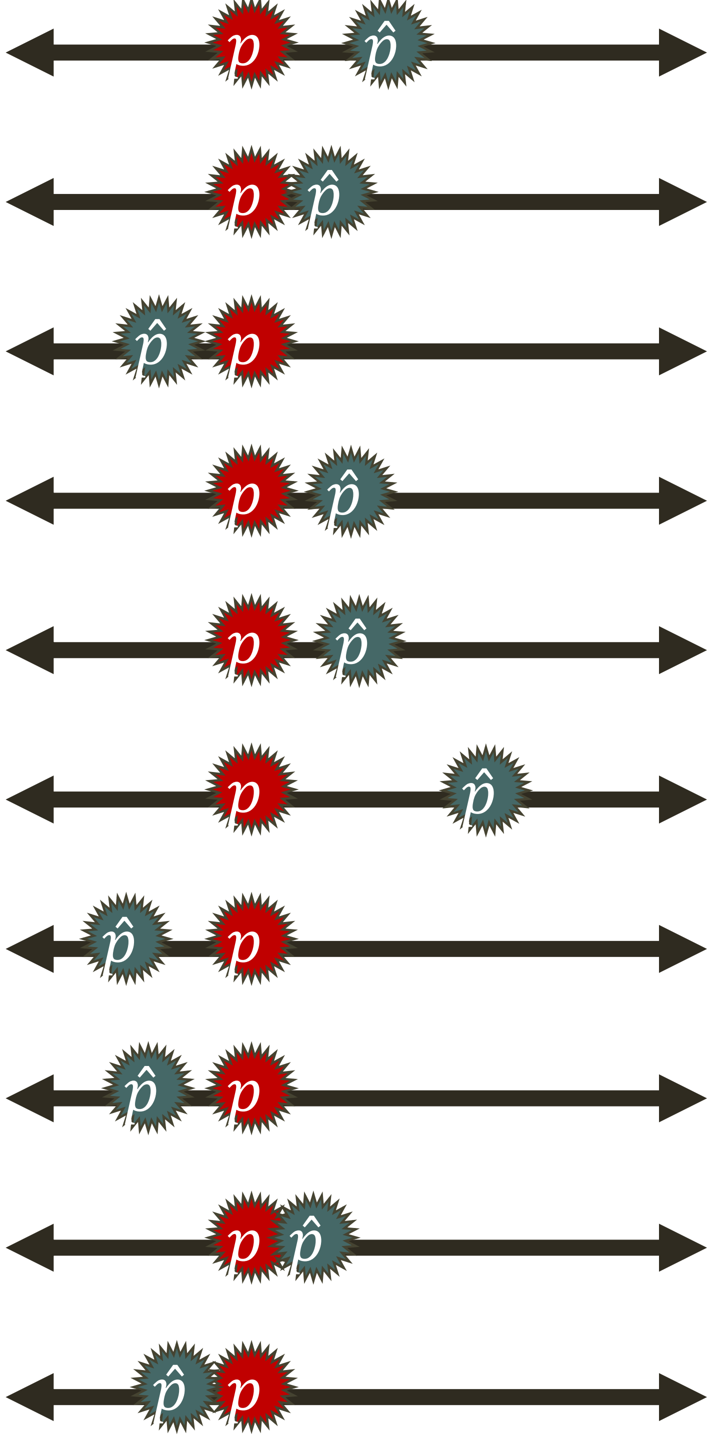
What we know:

- Our sample statistics are random variables
- Given samples of the same size (n) from the same population, they will vary
- The sampling distribution visualizes this variance for us

Implications:

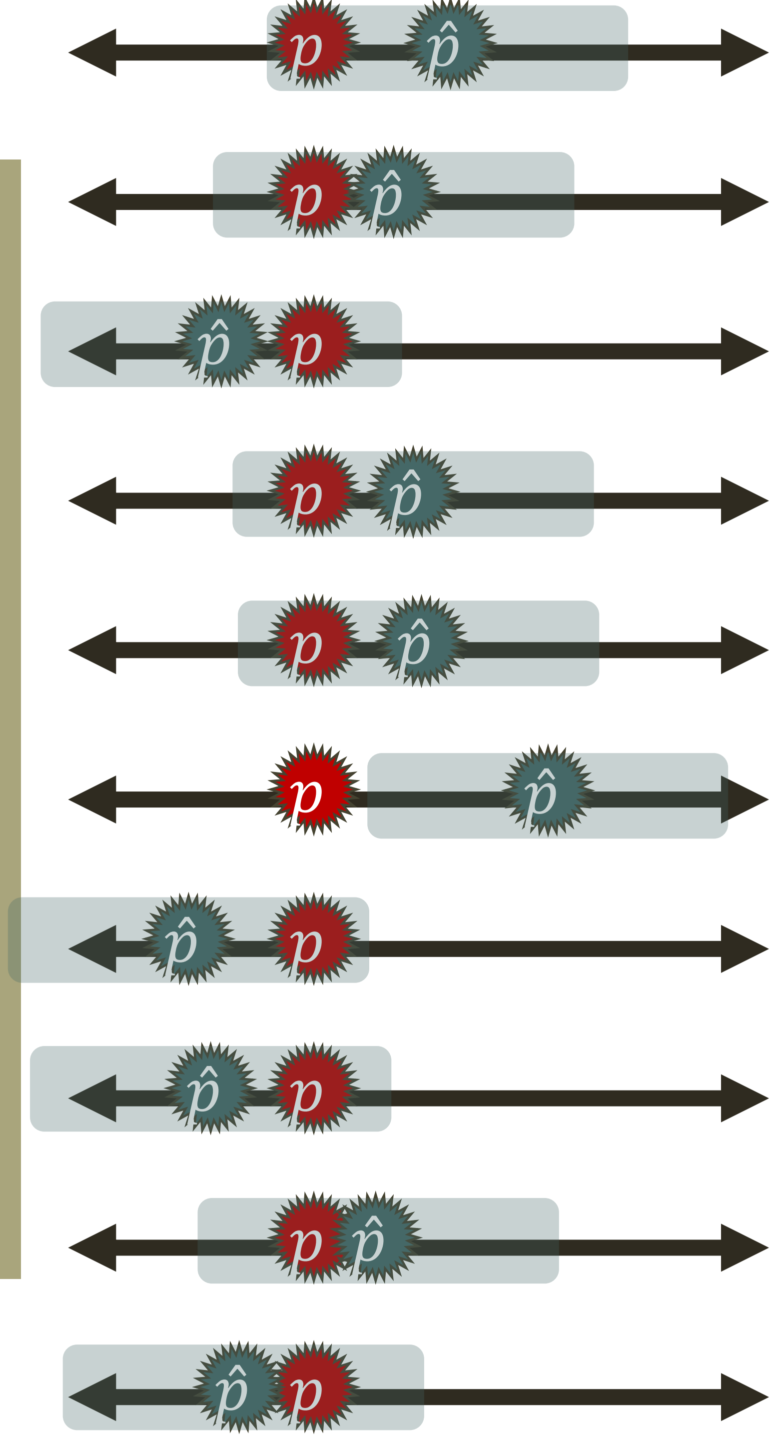
- While our sample statistics are our best guess for a true population parameter, they are not perfect
- Since we can't know the true population mean instead we might ask:
    - What are a range of plausible values for our population parameter (p) that our observed data and sample estimate ($\hat{p}$) are consistent with?

Confidence Intervals



With one estimate, I'm very unlikely to exactly hit the population parameter.

Confidence Intervals



With one estimate, I'm very unlikely to exactly hit the population parameter.

With a range of plausible values, I have a high chance of capturing the population parameter.

← Notice I only miss once

Confidence Intervals

A ***confidence interval*** is an interval providing a range of plausible values for our sample statistic, given the observed data.

We denote a confidence interval as (lower bound, upper bound), $p \pm range = (p - range, p + range)$

The ***confidence level*** of an interval represents the long run percent of intervals that capture the population parameter.

We denote confidence level $1 - \alpha$. Typically, $\alpha = 0.05$, so our confidence is 95%

Confidence Intervals

A ***confidence interval*** is an interval providing a range of plausible values for our sample statistic, given the observed data.

We denote a confidence interval as (lower bound, upper bound), $p \ \pm range = (p - range, p + range)$

The ***confidence level*** of an interval represents the long run percent of intervals that capture the population parameter.

We denote confidence level $1 - \alpha$. Typically, $\alpha = 0.05$, so our confidence is 95%

Important: A 95% confidence interval on a sample parameter means we are 95% confidence that our interval captures the true population parameter.

Remember, the *population parameter does not change*, our sample estimates are what vary.

Important: A 95% confidence interval on a sample parameter means we are 95% confidence that our interval captures the true population parameter.

Remember, the *population parameter does not change*, our sample estimates are what vary.

Confidence Intervals

Practice:

Let an experiment be rolling a loaded die (loaded to land on six 80% of the time) 1000 times and counting the proportion of sixes.

In this scenario, $p = 0.80$, and

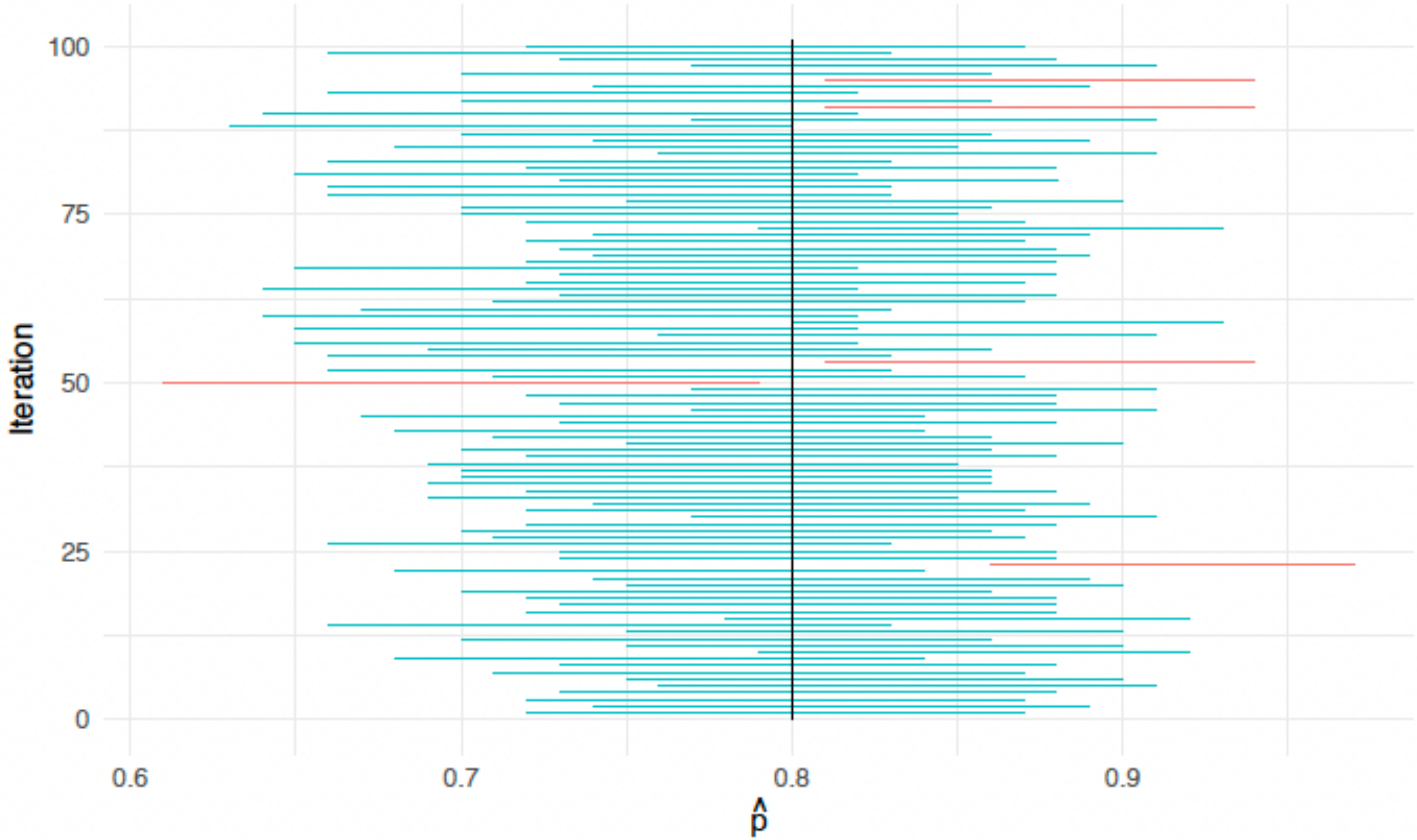
$\hat{p} = proportion\ of\ sixes\ rolled\ in\ the\ experiment$

If I repeat this experiment 100 times and construct a 95% confidence interval (CI) of $\hat{p}$ each time, how many CI's would you expect to capture p?

Confidence Intervals

Important: A 95% confidence interval on a sample parameter means we are 95% confidence that our interval captures the true population parameter.

Remember, the *population parameter does not change*, our sample estimates are what vary.



Important: A 95% confidence interval on a sample parameter means we are 95% confidence that our interval captures the true population parameter.

Remember, the *population parameter does not change*, our sample estimates are what vary.

Confidence Intervals

Practice:
Would you expect a 90% confidence interval to capture the population parameter more or less often than a 95% one?

What about an 80% confidence interval? A 99% one?

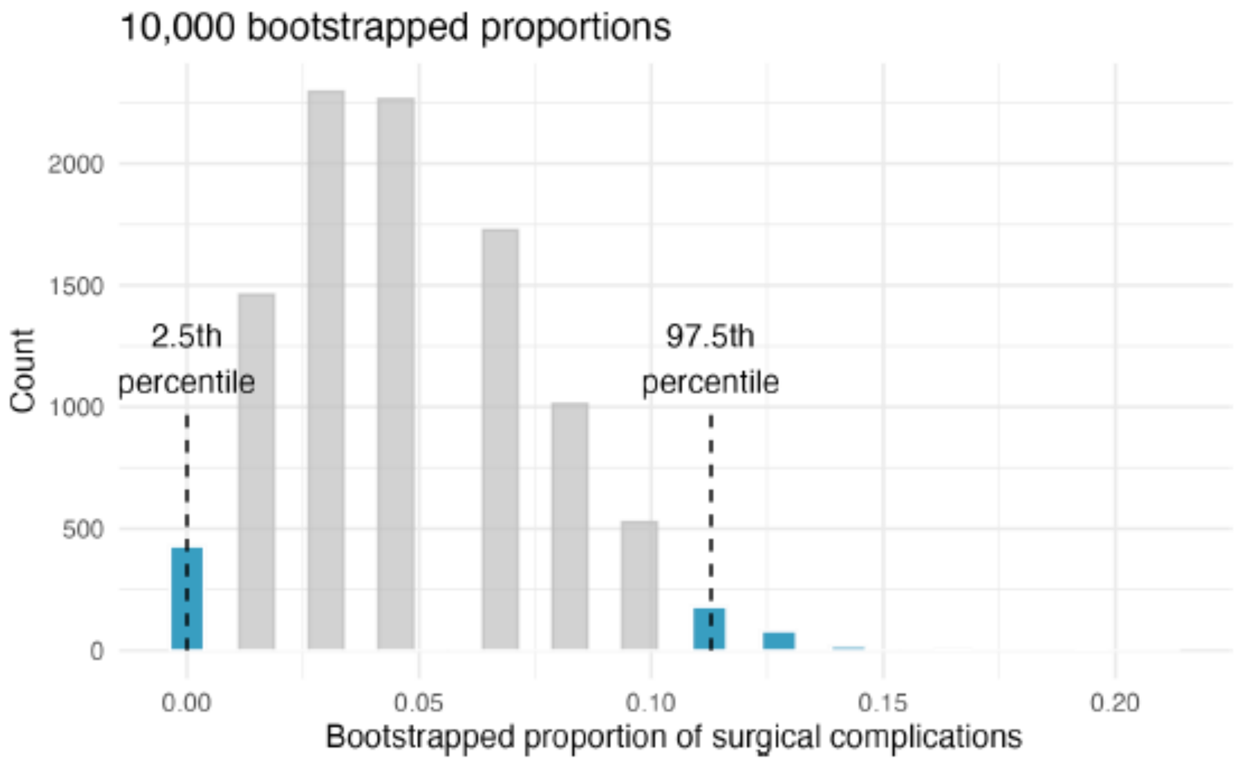
Would all of these intervals be equally useful to you?

$\hat{p}$

Confidence Intervals

Finding the 95% confidence interval for an estimate

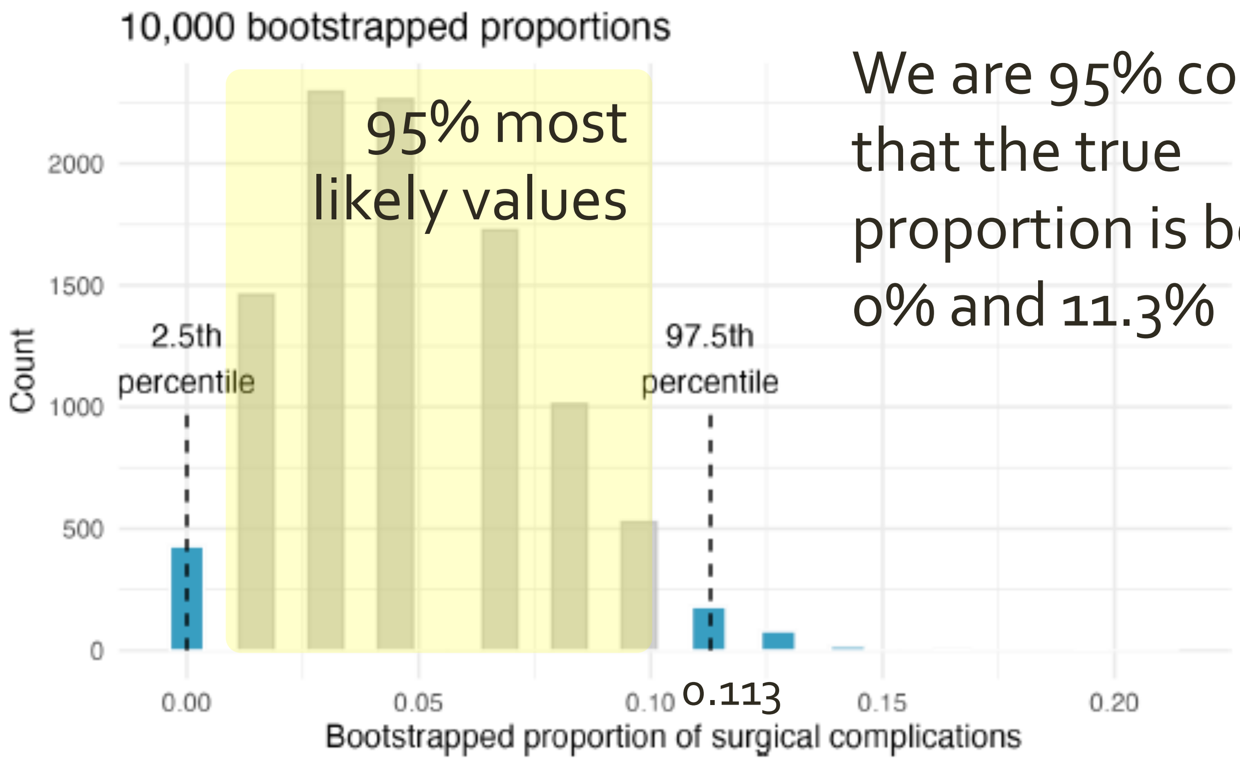
- Using the boostrapped distribution for our sample estimate, id the 2.5th percentile and 97.5th percentile of sample estimates
- These are less likely estimates that are far from the mean



Confidence Intervals

Finding the 95% confidence interval for an estimate

- Using the boostrapped distribution for our sample estimate, id the 2.5th percentile and 97.5th percentile of sample estimates
- These are less likely estimates that are far from the mean
- The remaining 95% of estimates are our confidence interval

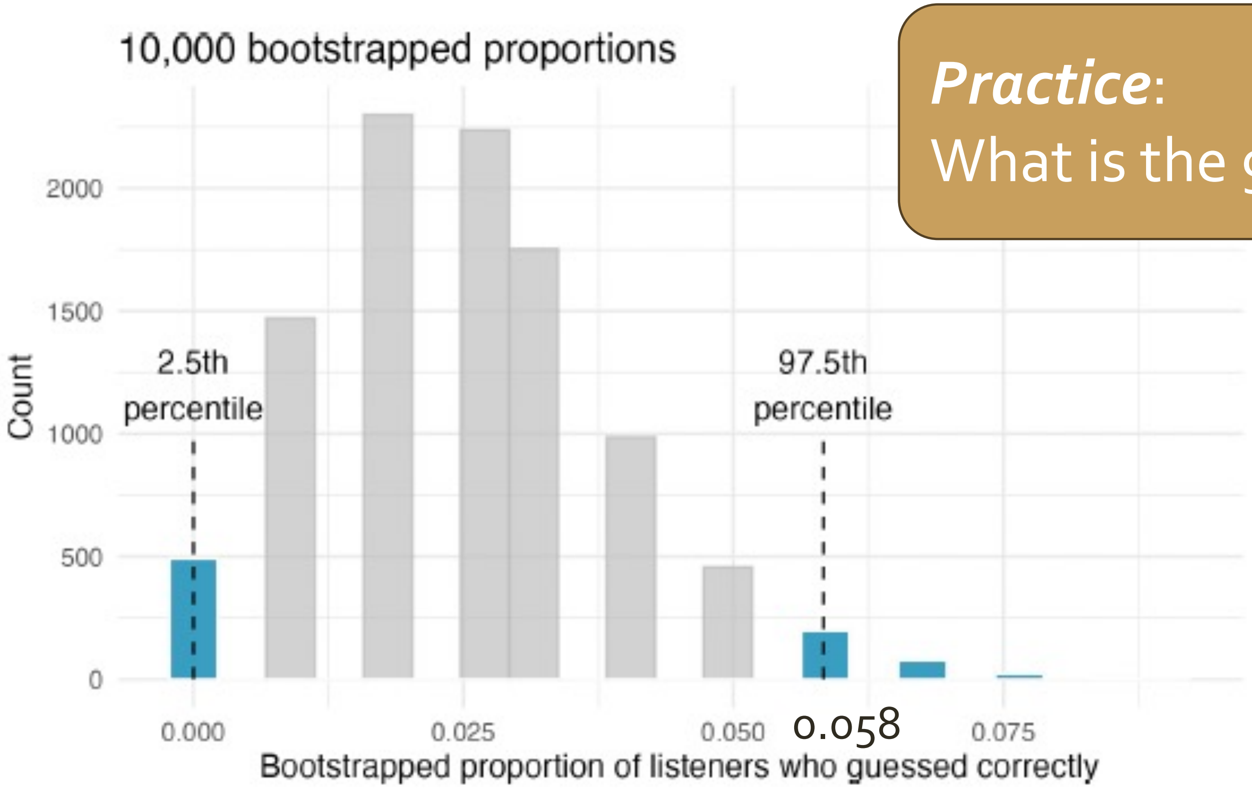


We are 95% confident that the true proportion is between 0% and 11.3%

Confidence Intervals

Finding the 95% confidence interval for an estimate

- Using the boostrapped distribution for our sample estimate, id the 2.5th percentile and 97.5th percentile of sample estimates
- These are less likely estimates that are far from the mean
- The remaining 95% of estimates are our confidence interval

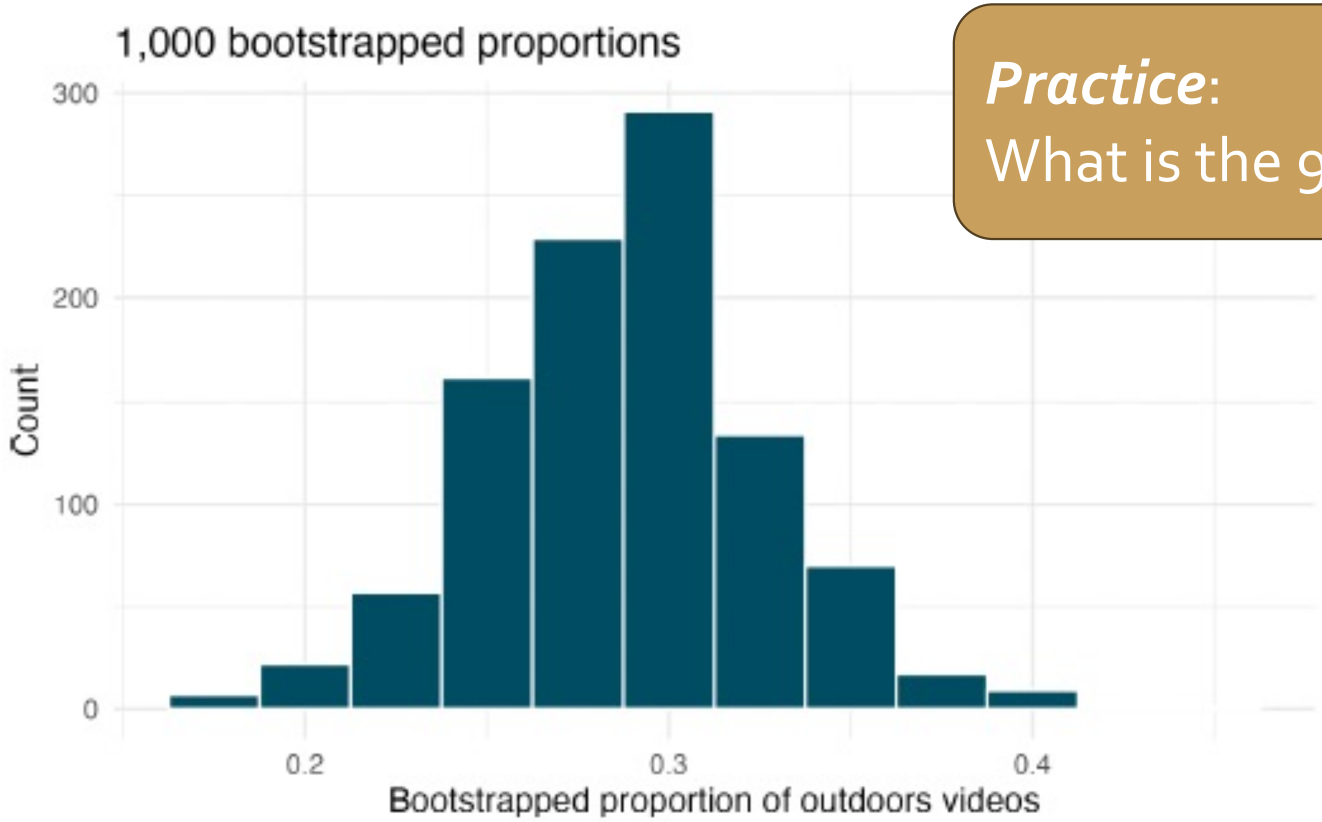


Practice:
What is the 95% CI?

Confidence Intervals

Finding the 95% confidence interval for an estimate

- Using the boostrapped distribution for our sample estimate, id the 2.5th percentile and 97.5th percentile of sample estimates
- These are less likely estimates that are far from the mean
- The remaining 95% of estimates are our confidence interval



Practice:
What is the 90% CI?

Confidence Intervals

Practice: Teens were surveyed about cyberbullying, and 54% to 64% reported experiencing cyberbullying (95% confidence interval). Answer the following questions based on this interval. (Pew Research Center, 2018)

a. A newspaper claims that a majority of teens have experienced cyberbullying. Is this claim supported by the confidence interval? Explain your reasoning.

b. A researcher conjectured that 70% of teens have experienced cyberbullying. Is this claim supported by the confidence interval? Explain your reasoning.

c. Without actually calculating the interval, determine if the claim of the researcher from part (b) would be supported based on a 90% confidence interval?