The one-sample bootstrap

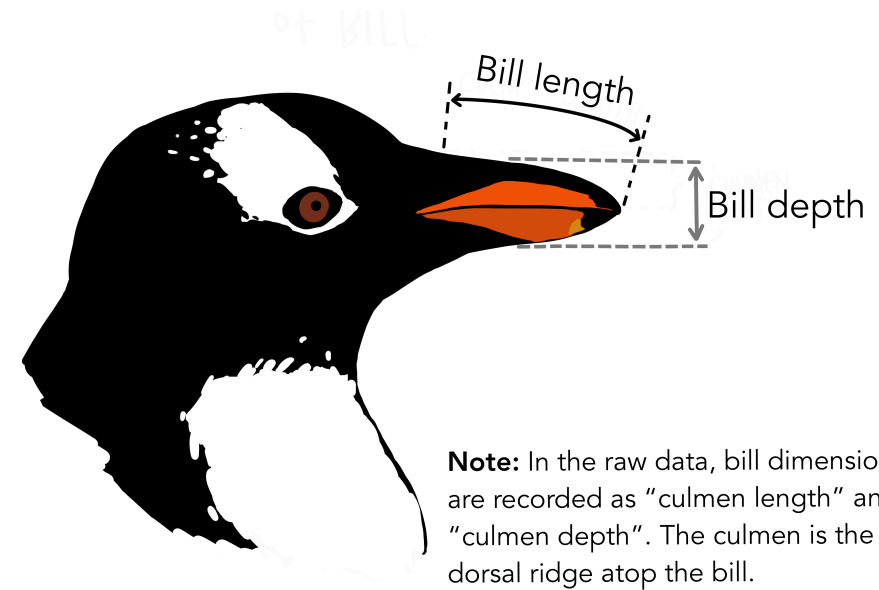Stat 250

Click here for PDF version

Roadmap

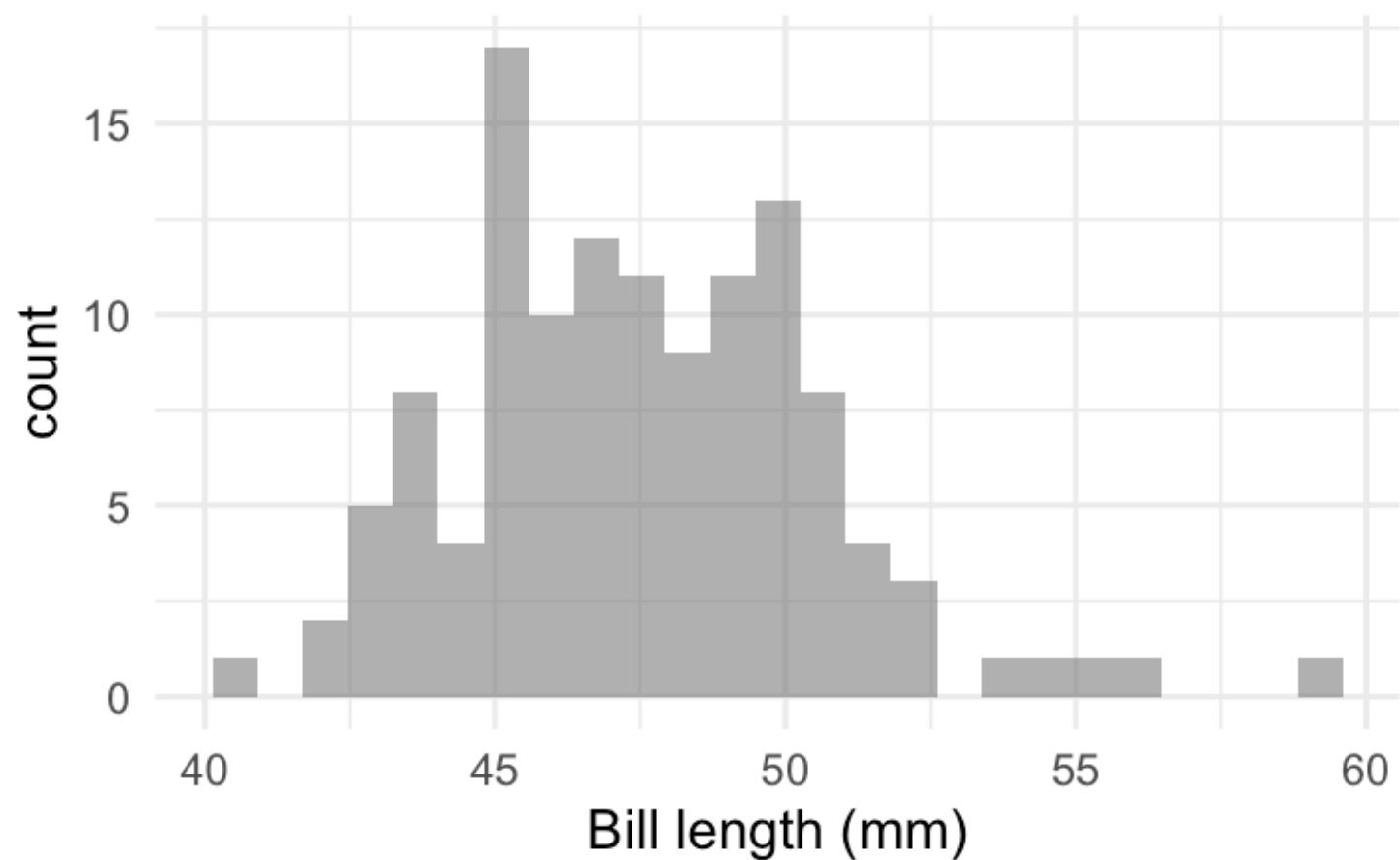We want to develop an **interval estimate** of a population parameter

1. *Exact method:* Find the sampling distribution in closed form (Chapter 4). Requires knowledge of the distribution of the data!

2. **Bootstrap method:** Use the sample to approximate the population and simulate a sampling distribution (Chapter 5).

3. *Asymptotic method:* Use large-sample theory to approximate the sampling distribution (e.g., appeal to CLT; Chapter 7)

Example

- Dr. Kristen Gorman and the Palmer Station, Antarctica LTER, are studying the bill dimensions of a certain species of penguin

- They want to estimate the average bill depth and bill length (in mm)

Note: In the raw data, bill dimensions are recorded as "culmen length" and "culmen depth". The culmen is the dorsal ridge atop the bill.

Bill length

min	Q1	median	Q3	max	mean	sd
40.9	45.3	47.3	49.55	59.6	47.5	3.08

How can we use our one sample to estimate the population mean bill length for all Gentoo penguins?

Bootstrap distribution
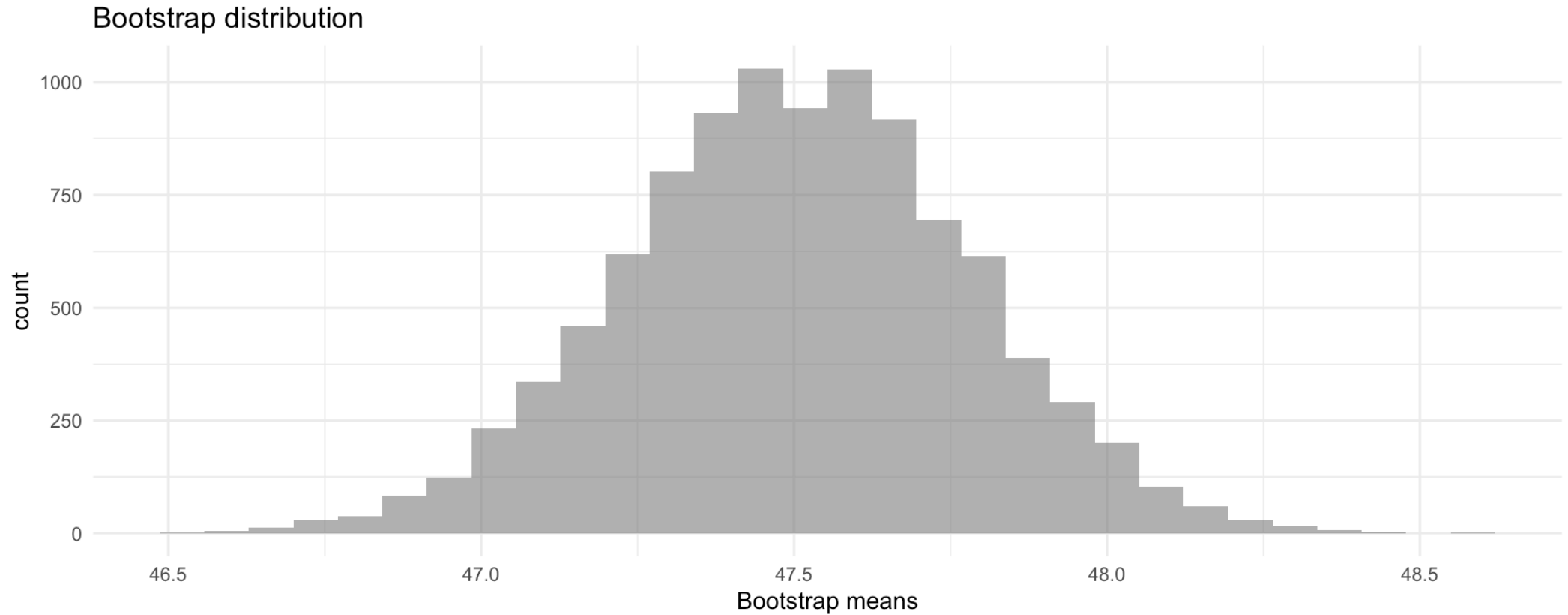
Bootstrap percentile interval

A 95% confidence interval can be constructed from the 2.5 and 97.5th percentiles of the bootstrap distribution

The one-sample bootstrap algorithm

Given a sample of size *n* from a population,

1. Draw a resample of size *n*, **with replacement**, from the sample.

2. Compute the statistic of interest.

3. Repeat this resampling process (steps 1-2) many times, say 10,000.

4. Construct the bootstrap distribution of the statistic.

Your turn

A sample consists of the following values: 8, 4, 11, 3, 7.

Which of the following are possible bootstrap samples from this sample? Why?
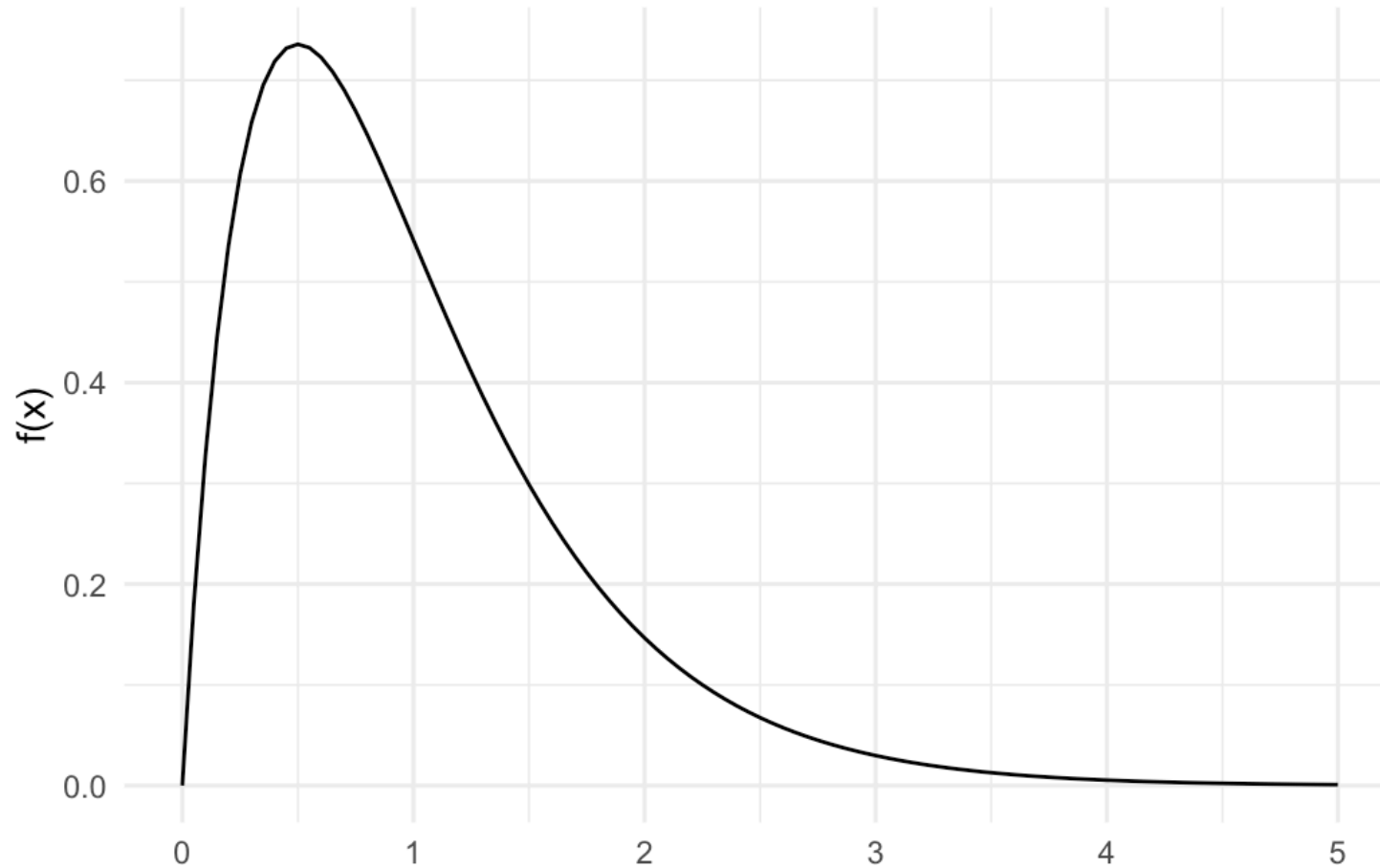
a. 8, 3, 7, 11

b. 4, 11, 4, 3, 3

c. 3, 4, 5, 7, 8

d. 7, 8, 8, 3, 4

How does the bootstrap work?

Population

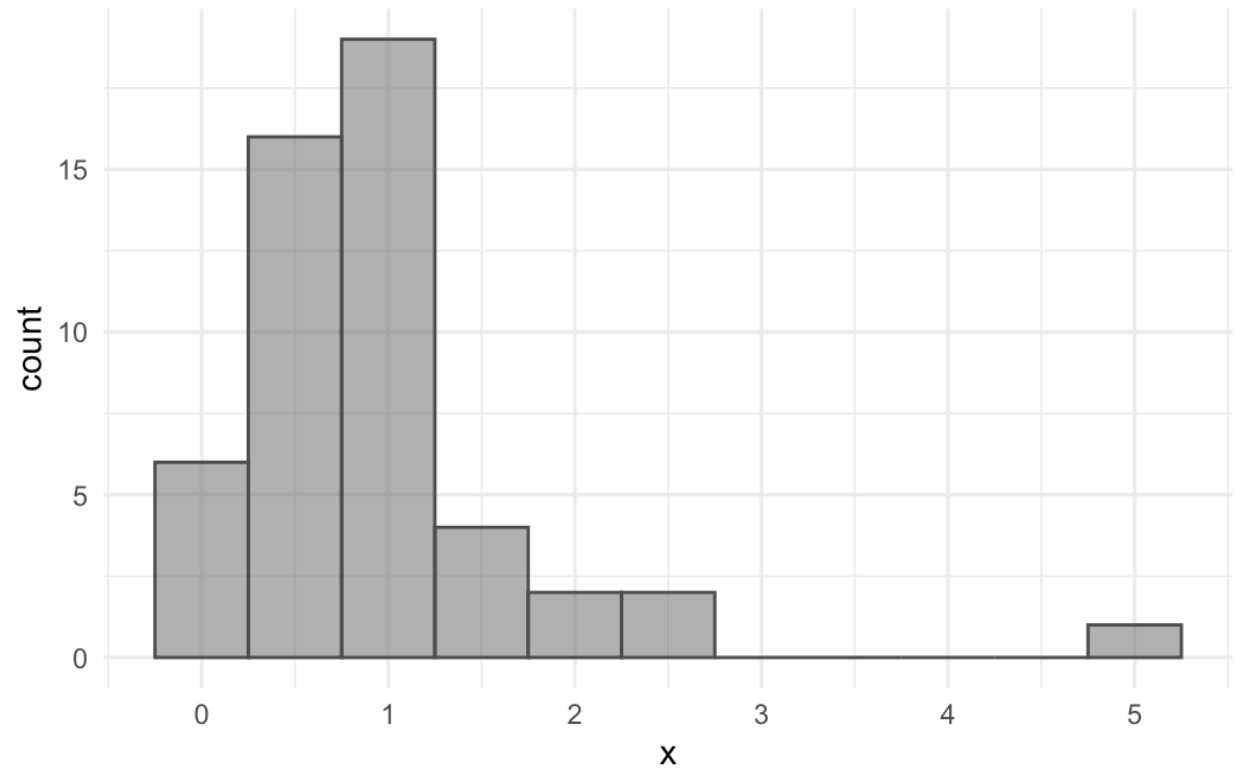Consider a Gamma(2, 2) population distribution
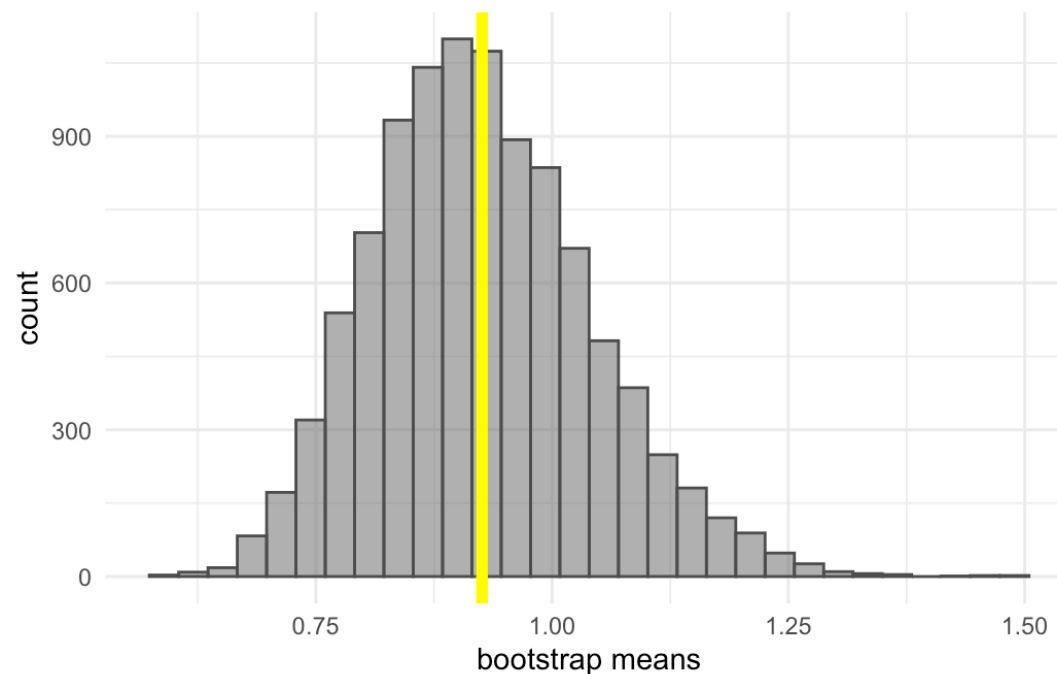

$$E(X) = 1 \qquad SD(X) = 1/2$$

Sample

Suppose we draw a random sample of size $n = 50$

min	Q1	median	Q3	max	mean	sd	
0.06	0.45	0.76	1.07	5.09	0.92	0.82	5

Bootstrap distribution

We can bootstrap our sample and obtain the bootstrap distribution

min	Q1	median	Q3	max	mean	sd	
0.59	0.84	0.92	1	1.49	0.93	0.12	10(

Sampling distribution

We could also draw many different samples and obtain the sampling distribution
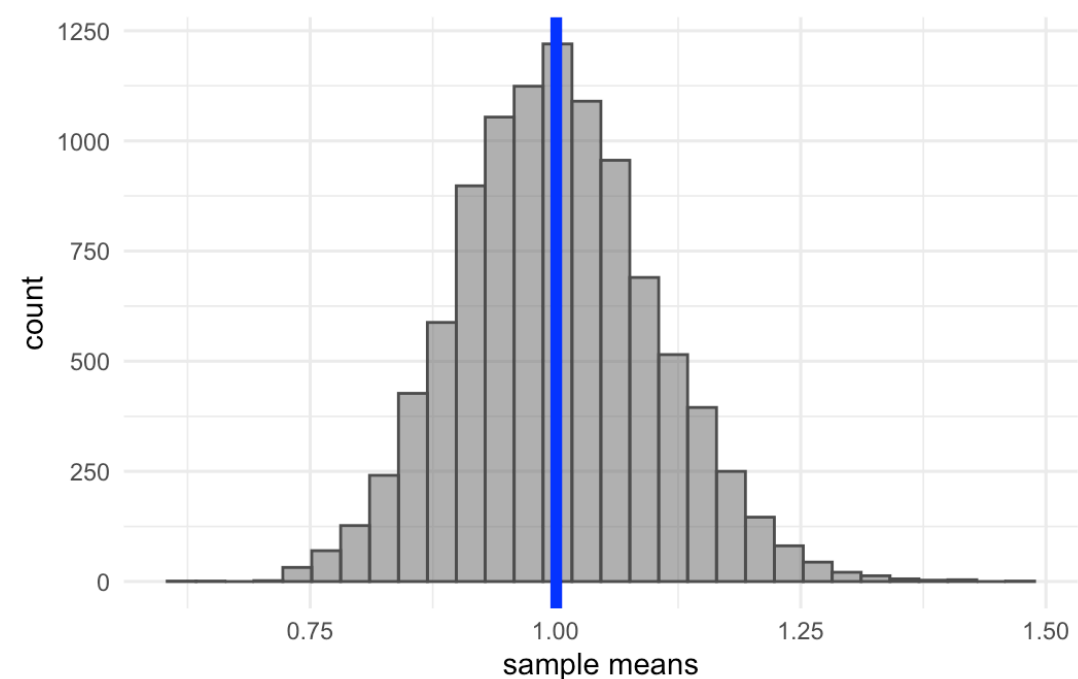
min	Q1	median	Q3	max	mean	sd	
0.6	0.93	1	1.07	1.46	1	0.1	100

Key comparisons

	Mean	SD	Bias
Population	1	0.5	
Sample	0.92	0.925	
Sampling distribution	1	0.1	0.001
Bootstrap distribution	0.93	0.115	0.001

![R] implementation

```r
# Subsetting to get only one species
gentoo <- dplyr::filter(penguins, species == "Gentoo")

# Bookkeeping
y <- gentoo$bill_length_mm
n <- nrow(gentoo)          # sample size
N <- 10^4                  # desired no. resamples
boot_means <- numeric(N) # a place to store the bootstrap stats

# Resampling from the sample
for (i in 1:N) {
  x <- sample(y, size = n, replace = TRUE)
  boot_means[i] <- mean(x, na.rm = TRUE)  # you can choose other stati:
}
# Calculate a 95% percentile interval
quantile(boot_means, probs = c(0.025, 0.975))
```