Bootstrap

- *Non-parametric Bootstrap*
- *Parametric Bootstrap*

James Balamuta

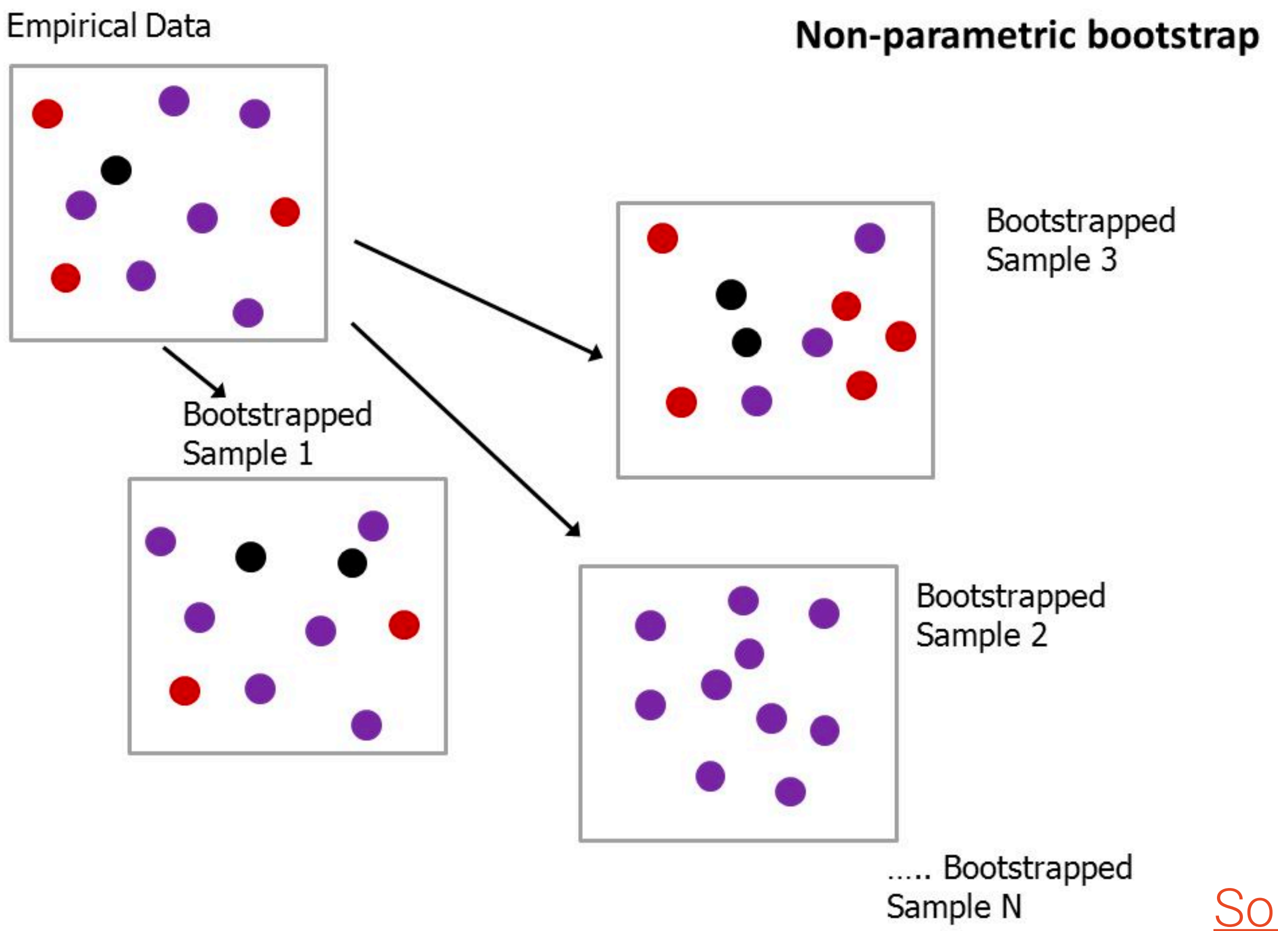
IMSI Data Science Workshop Series Summer 2022

Lecture Objectives

- **Understanding scenarios** to use bootstrap
- **Importance of resampling** in the bootstrap procedure.
- **Differences** between **parametric** and **non-parametric** bootstrap.
- **Creating** and **interpreting quantile** confidence intervals.

Non-parametric Bootstrap

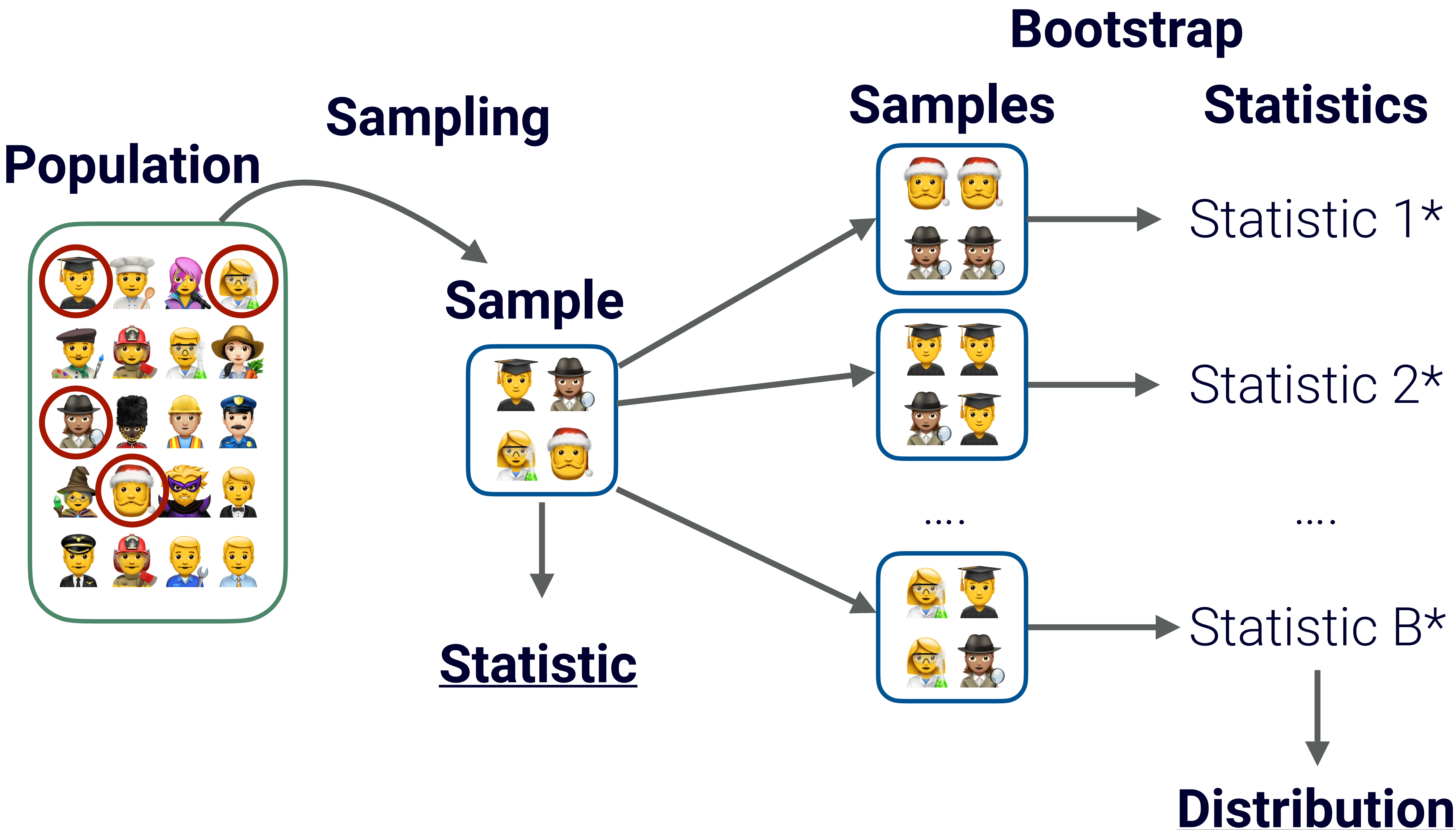
Definition:
Non-parametric Bootstrap seeks to estimate an underlying probability distribution by resampling observed values.



Source

Bootstrap Idea

... in a scene ...



In Action

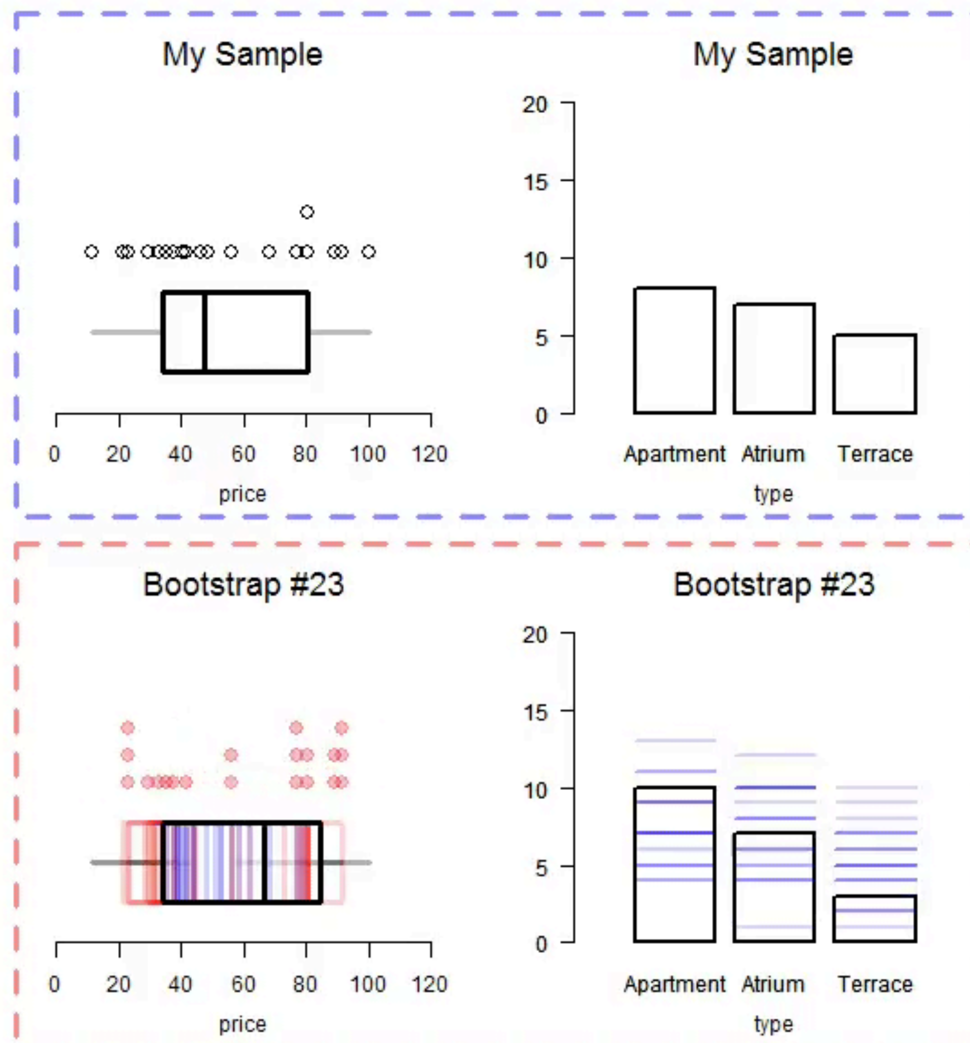
… what's happening …

My Sample

price	type
40.338	Terrace
68.531	Apartment
21.551	Atrium
46.214	Terrace
32.983	Atrium
80.411	Apartment
80.607	Apartment
77.210	Apartment
23.318	Atrium
55.891	Terrace
80.568	Apartment
41.809	Terrace
11.344	Atrium
89.045	Apartment
100.063	Apartment
35.037	Atrium
29.517	Atrium
37.408	Atrium
48.840	Terrace
91.330	Apartment

Bootstrap #23

price	type
91.330	Apartment
89.045	Apartment
77.210	Apartment
77.210	Apartment
35.037	Atrium
55.891	Terrace
80.607	Apartment
32.983	Atrium
23.318	Atrium
23.318	Atrium
55.891	Terrace
29.517	Atrium
91.330	Apartment
37.408	Atrium
23.318	Atrium
41.809	Terrace
91.330	Apartment
77.210	Apartment
80.568	Apartment
89.045	Apartment



Source

Why Bootstrap?

… where's the infinite data ???

1. Sampling data takes both **time** and **money**

2. No ability to make an assumption about the sample's underlying distribution…

3. Inference done with asymptotic theory on samples may not make sense **if values have a restriction**.
(e.g. height cannot be negative or zero)

Bootstrap Terms

... describing non-parametric bootstrap statistically ...

Real World

Unknown Distribution

Observed Values

$$F \to X = \left(X_1, \cdots, X_{\boxed{n}}\right)$$

$$\hat{\theta} = s\left(X\right)$$

Statistic on **Observed** Values

Bootstrap World

Empirical Distribution

Bootstrap Sample

$$\hat{F}_n \to X^{*,i} = \left(X_1^{*,i}, \cdots, X_{\boxed{n}}^{*,i}\right)$$

$$\hat{\theta}^{*,i} = s\left(X^{*,i}\right)$$

Statistic on **Bootstrapped** Values

$\hat{\theta}^*$ **to** $\hat{\theta}$ **is like** $\hat{\theta}$ **to** θ

Statistic on
Bootstrapped Values

Statistic on
Observed Values

Population Statistic

Resampling

generating fictional data from real data ...

Original Data

$$F \rightarrow X = \left(X_1, \cdots, X_n\right)$$

	id	sex	height
X_1	1	M	6.1
X_2	2	F	5.5
X_3	3	F	5.2
X_4	4	M	5.6
X_5	5	M	5.9

X_1^*

X_n^*

Resampled Data

$$\hat{F}_n \rightarrow X^* = \left(X_1^*, \cdots, X_n^*\right)$$

	id	sex	height
$X_1^{*,1}$	2	F	5.5
$X_2^{*,1}$	3	F	5.2
$X_3^{*,1}$	3	F	5.2
$X_4^{*,1}$	1	M	6.1
$X_5^{*,1}$	4	M	5.6

...

...

	id	sex	height
$X_1^{*,n}$	4	M	5.6
$X_2^{*,n}$	4	M	5.6
$X_3^{*,n}$	2	F	5.5
$X_4^{*,n}$	1	M	6.1
$X_5^{*,n}$	2	F	5.5

New Samples

Sample **with** Replacement

Your Turn

Consider the following **sample** taken from a population:

Red, **Green**, **Blue**, **Orange**

Would the following be considered a **valid bootstrap sample**?

Red, **Green**, **Green**, **Purple**

Strategy

… how to roll a **non-parametric** bootstrap …

Step 1: Obtain a sample of the population

Step 2: Using resampling technique, construct

$$\hat{F}_n \overset{iid}{\sim} X^{*,i} = \left(X_1^{*,i}, X_2^{*,i}, \cdots, X_n^{*,i} \right)$$

that contains the same number of observations as

$$F \overset{iid}{\sim} X = \left(X_1, X_2, \cdots, X_n \right)$$

Step 3: Compute a statistic on the resampled data as $\hat{\theta}^* = s\left(X^{*,i} \right)$

Step 4: Repeat **Steps 2 - 3** until i matches required number of iterations.

Overall Procedure

Statistic

$\hat{\theta}^*$

Resampled Data

$$\hat{F}_n \rightarrow X^* = \left(X_1^*, \cdots, X_n^*\right)$$

	id	sex	height
$X_1^{*,1}$	2	F	5.5
$X_2^{*,1}$	3	F	5.2
$X_3^{*,1}$	3	F	5.2
$X_4^{*,1}$	1	M	6.1
$X_5^{*,1}$	4	M	5.6

$\rightarrow \hat{\theta}^{*,1}$

Original Data

$$F \rightarrow X = \left(X_1, \cdots, X_n\right)$$

	id	sex	height
X_1	1	M	6.1
X_2	2	F	5.5
X_3	3	F	5.2
X_4	4	M	5.6
X_5	5	M	5.9

X_1^*

X_n^*

…

…

…

…

	id	sex	height
$X_1^{*,n}$	4	M	5.6
$X_2^{*,n}$	4	M	5.6
$X_3^{*,n}$	2	F	5.5
$X_4^{*,n}$	1	M	6.1
$X_5^{*,n}$	2	F	5.5

$\rightarrow \hat{\theta}^{*,n}$

Sample **with** Replacement

Your Turn

Consider a sample of **size 40** (*n = 40*)…

If we generated **10,000 bootstrap** replications:

- How many observations will be present for *each* bootstrap replication?
- What is the number of bootstrap statistics that will be created?

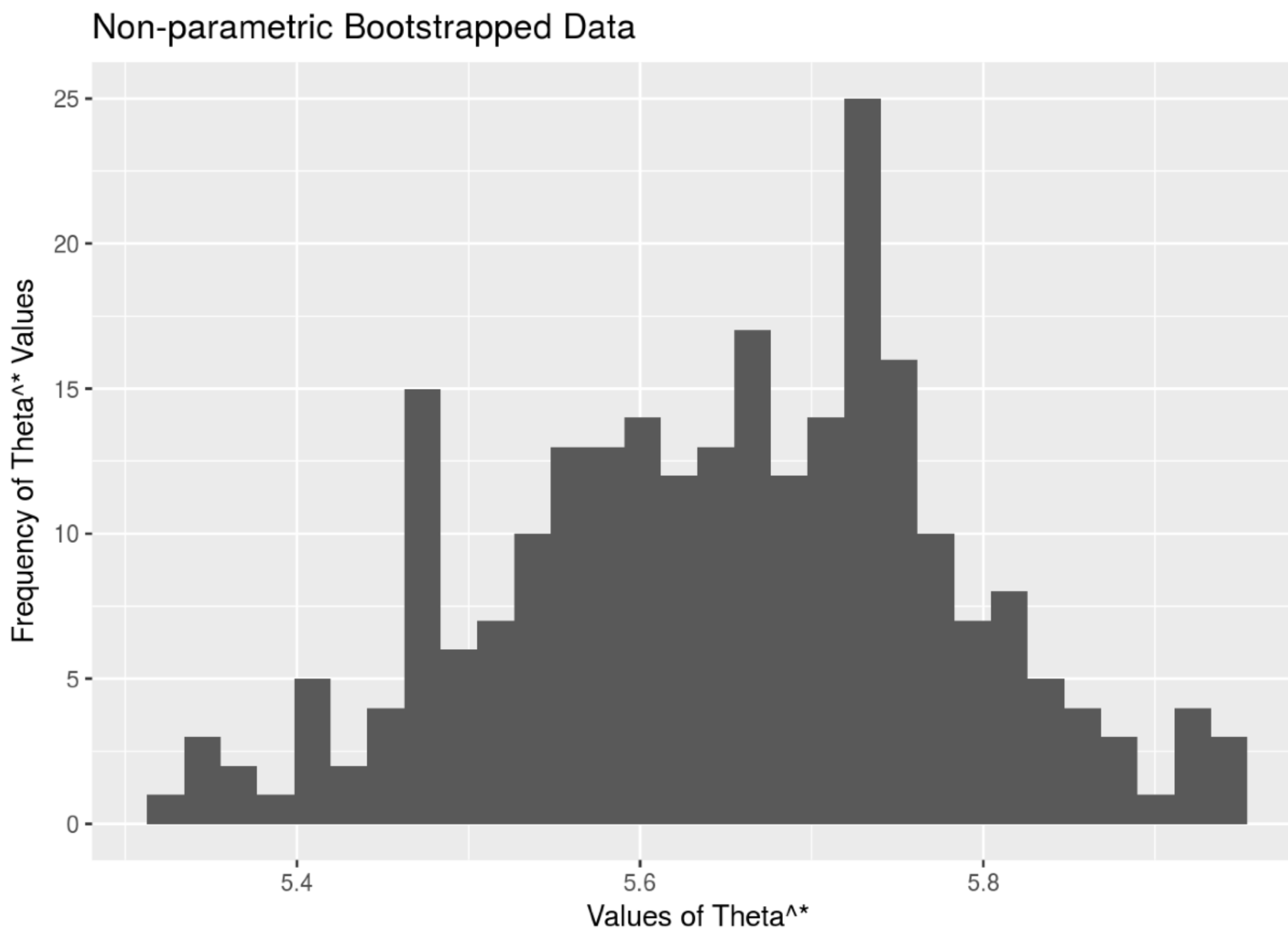
Implementation

... of **non**-parametric bootstrap ...

```
sample_data = ???                        # Step 1: Obtain samples from population
theta_hat = mean(sample_data$var)        # Compute the mean for the data
n_obs = nrow(sample_data)                # Length of data
boot_iter = 250L                         # Number of bootstrap iterations
theta_star = rep(NA, boot_iter)          # Bootstrapped estimate of theta

for (i in seq_len(boot_iter)) {
  set.seed(11882 + i)                    # Set seed for reproducibility
  # Step 2: Randomly sample observations positions from 1 to n_obs
  indexes = sample(n_obs, n_obs, replace = TRUE)

  # Extract out the observation positions
  sample_data_star = sample_data[indexes,, drop = FALSE]

  # Step 3: Compute the desired statistic on the bootstrapped values
  theta_star[i] = mean(sample_data_star$var)
} # Step 4: Repeat until i matches boot_iter
```

Bootstrapped Distribution

… values of bootstrapped statistic …



Your Turn

Modify the bootstrap so that it computes the **standard deviation** of **Sepal.Width** in the **iris** data set.

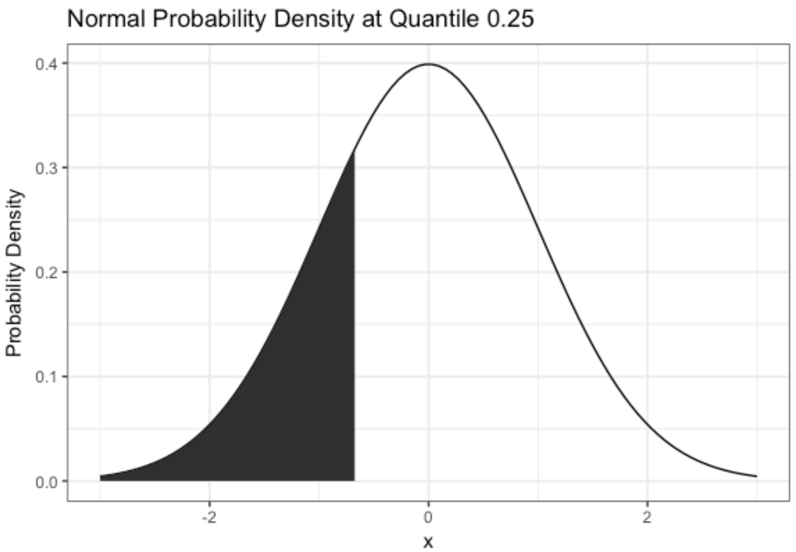
Recall: The **sd()** function provides the standard deviation

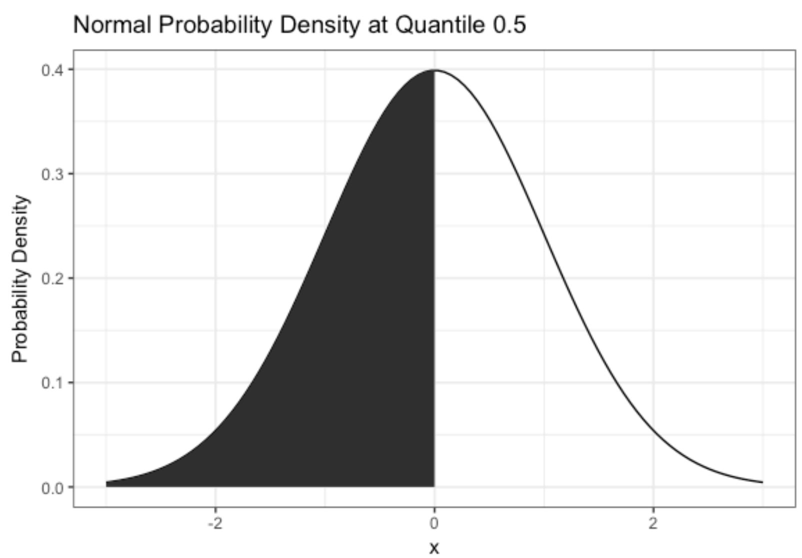
$$\sigma = sd\left(x\right) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}$$

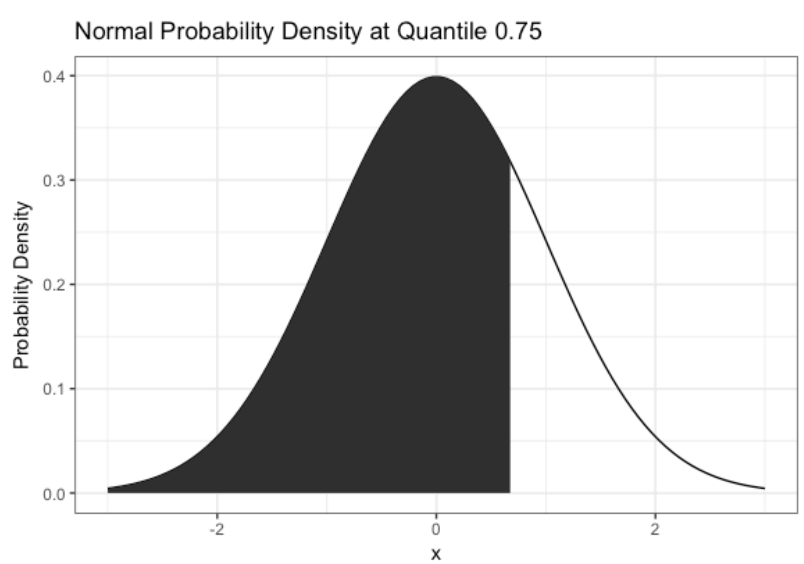
$$\overline{x} = mean\left(x\right) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Quantiles / Percentiles

… value of the distribution at *p*-th location …

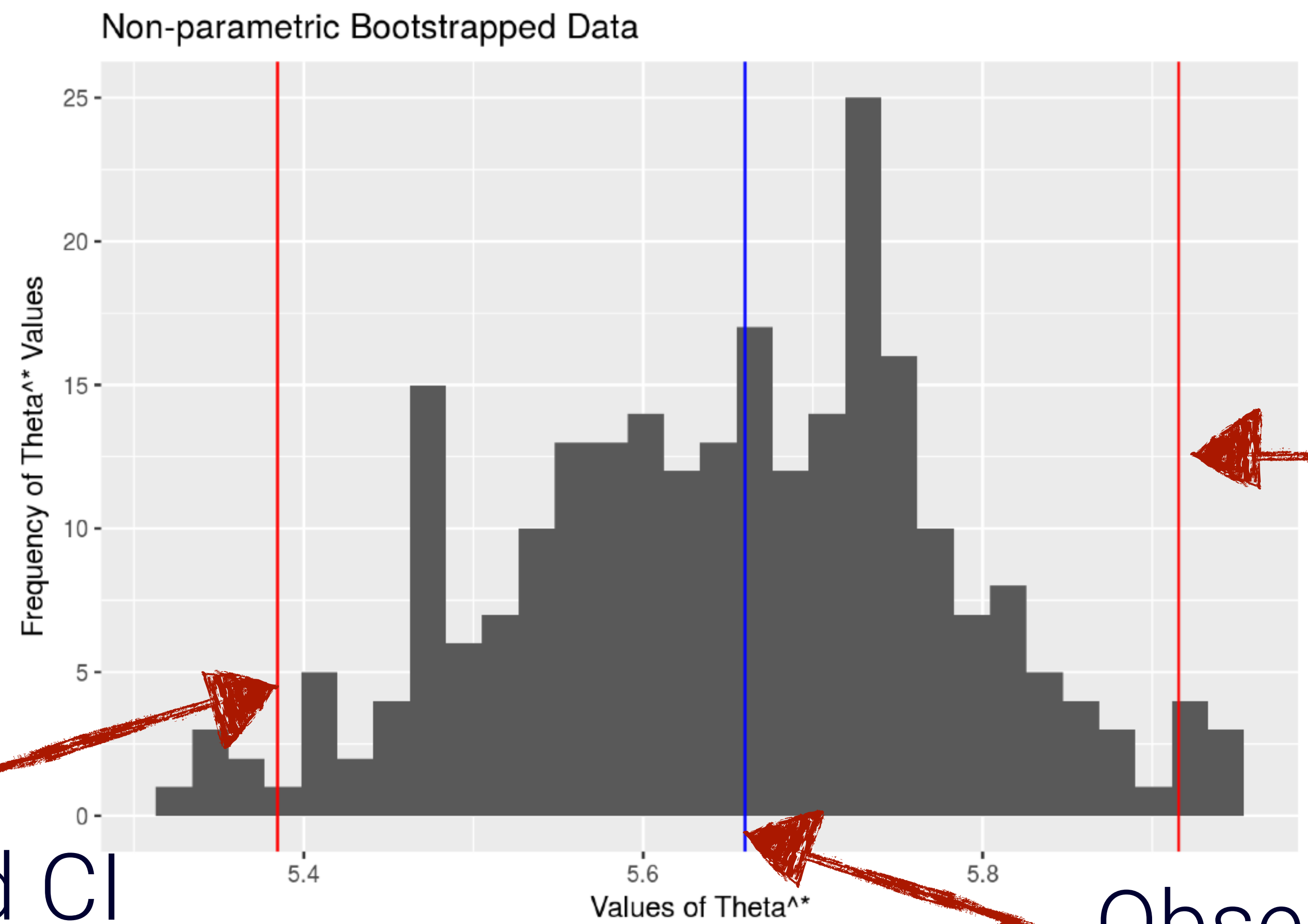







```r
# Sample Data
x = c(1, 2, 3, 4, 5, 6)

# Value at ordered point in
# distribution
quantile(x,
  probs = c(0.25, 0.5, 0.75, 1)
)
#  25%   50%   75%  100%
# 2.25  3.50  4.75  6.00

# Median is the 50% quantile
median(x)
# [1] 3.5
```

```r
# Percentile CIs
# Computes a custom confidence interval with data...

# Significance level
alpha = 0.05

# Under alpha = 0.05, we are retrieving quantiles for 0.025 and 0.975
ci_range = quantile(theta_star, probs = c(alpha / 2, 1 - alpha / 2))
# [1] low high
```



Upper Bootstrapped CI

Lower Bootstrapped CI

Observed Sample Mean

Your Turn

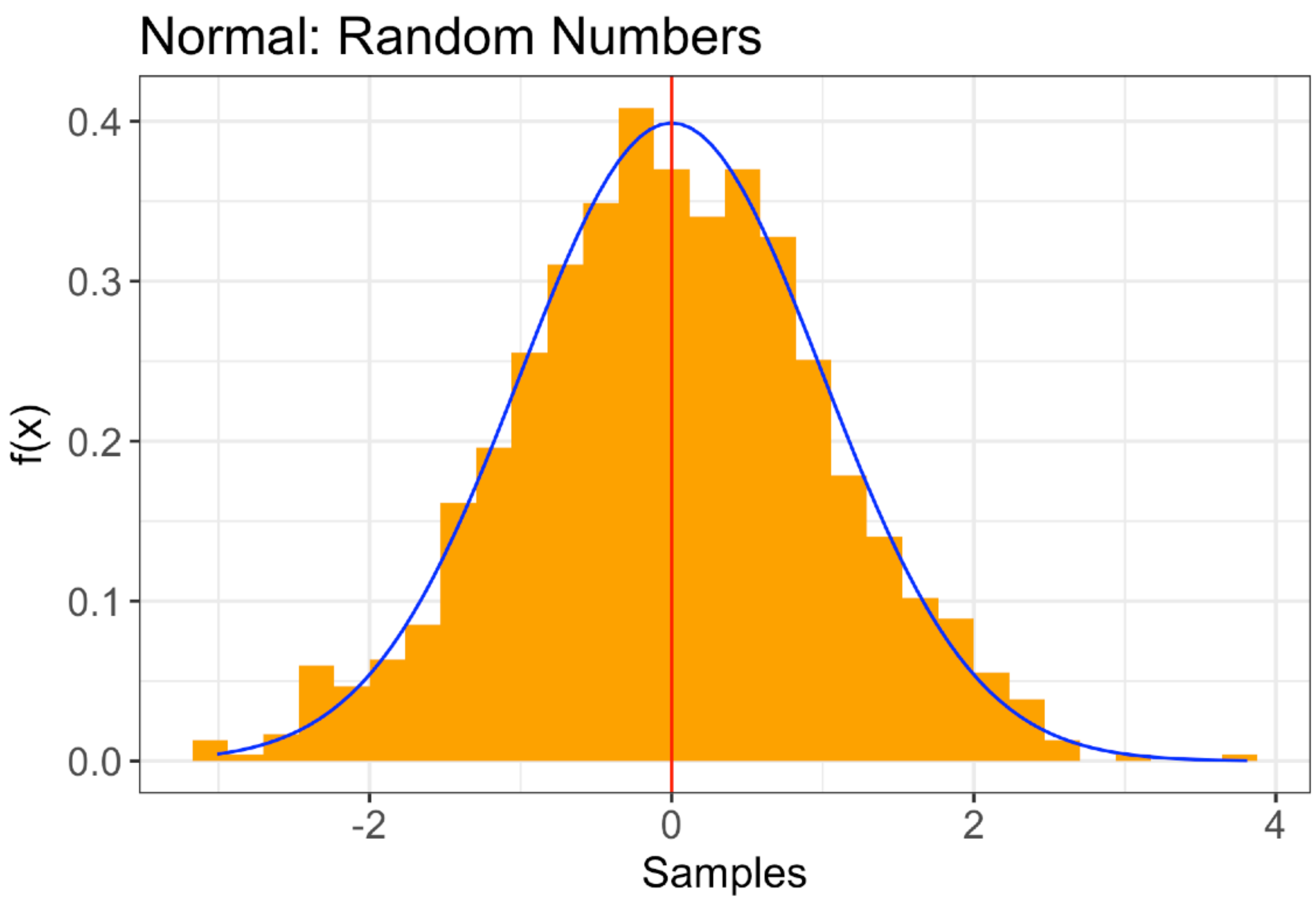
How does changing the **alpha value / significance level** used to construct the percentile confidence interval affect the length of the interval?

Consider two cases:

- 95% -> 90%
- 95% -> 97%

Parametric Bootstrap

Definition:
Parametric Bootstrap seeks to estimate parameters under the assumption they belong to a specific family of probability distributions.



Strategy

… how to roll a **parametric** bootstrap …

Step 1: Sample model data under a known distribution with unknown parameters θ.

$$F_{\theta} \overset{iid}{\sim} X = \left(X_1, X_2, \cdots, X_n\right)$$

Step 2: Compute the statistic under the known distribution $\hat{\theta} = s\left(X\right)$

Step 3: Sample model under $F_{\hat{\theta}} \overset{iid}{\sim} X^{*,i} = \left(X_1^{*,i}, X_2^{*,i}, \cdots, X_n^{*,i}\right)$ via $\hat{\theta} = s\left(X\right)$

Step 4: Calculate the bootstrapped statistic $\hat{\theta}^{*,i} = s\left(X^{*,i}\right)$

Step 5: Repeat **Steps 3 - 4** until i matches required number of iterations.

Implementation

... of **parametric** bootstrap ...

```
sample_values = rnorm(100)                  # Step 1: Obtain samples from known
                                            # population distribution.



                                            # Step 2: Obtain statistics
theta_mean_hat = mean(sample_values)        # Compute sample mean
theta_sd_hat = sd(sample_values)            # Compute sample standard deviation

n_obs = length(sample_values)               # Length of data
boot_iter = 250L                            # Number of bootstrap iterations
theta_mean_star = rep(NA, boot_iter)        # Bootstrapped estimate of mean
theta_sd_star = rep(NA, boot_iter)          # Bootstrapped estimate of standard dev
```

Implementation

... of **parametric** bootstrap ...

```
# See previous slide for setup details…

for (i in seq_len(boot_iter)) {
 set.seed(385 + i)                          # Set seed for reproducibility

 # Step 3: Randomly generate observations under distribution
 sample_values_star = rnorm(n_obs, mean = theta_mean_hat, sd = theta_sd_hat )

 # Step 4: Compute the desired statistic on the bootstrapped values
 theta_mean_star[i] = mean(sample_values_star )
 theta_sd_star[i] = sd(sample_values_star )

} # Step 5: Repeat until i matches boot_iter
```

Parametric vs. Nonparametric

… what's the difference ???

Type	**Distribution**	**Parameter**
Non-parametric	*Unknown*	Unknown
Parametric	*Known*	Unknown

Redux

... highlighting the **difference** ...

```
## Nonparametric
# Unknown distribution. Sample values from observed distribution
indexes = sample(n_obs, n_obs, replace = TRUE)
# Extract out the observation positions
sample_data_star = sample_data[indexes,, drop = FALSE]

## Parametric
# Known distribution. Sample values underneath estimated parameters.
sample_values_star = rnorm(n_obs, mean = theta_mean_hat, sd = theta_sd_hat)
```

Your Turn

Implement a **parametric** bootstrap that determines the **mean**, **standard deviation**, and **median** of a Poisson distribution with

$$\lambda = 3$$

(lambda)

Recap

- **Non-parametric Bootstrap**
  - Resampling from a sample of the population to obtain an empirical distribution.
- **Parametric Bootstrap**
  - Sampling under a known probability distribution with estimated values of the initial sample.