Week 8 - Bootstrap Methods

STA238 - Winter 2025

Week of Feb. 24, 2025

In week 6, you were introduced to:

- The bootstrap principle
- Empirical bootstrap
- Parametric bootstrap
- Centred bootstrapped values

Bootstrap Principle

1. Use the data set $x_1, x_2, ..., x_n$ to compute an estimate $\hat{F}$ for the "true" distribution function F.
2. Mimic the random sample $X_1, X_2, ..., X_n$ from F with a random sample $X_1^*, X_2^2, ..., X_n^*$ from $\hat{F}$. Compute $h(x_1*, x_2^*, ..., x_n^*)$ for each bootstrap random sample.
3. Approximate the distribution of $h(X_1, X_2, ..., X_n)$ by that of $h(X_1^2, X_2^*, ..., X_n^*)$ by repeating steps (1) and (2) until you get B bootstrap samples.

Note:

- $x_1*, x_2^*, ..., x_n^*$ is one **bootstrap sample**.
- $\hat{\theta}^* = h(x_1*, x_2^*, ..., x_n^*)$ is one **bootstrapped point estimate**.

GOAL: Investigate how the estimator $\hat{\theta}$ behaves relative to the parameter of interest θ.

- In simulation: we can approximate distribution of $\hat{\theta}$ directly by sampling from a pre-defined model distribution F_θ.
- In bootstrapping: we investigate the behaviour constrained to estimated distribution $\hat{F}$. We should therefore compare how $\hat{\theta}^*$ behaves relative to the $\hat{\theta}$ in $\hat{F}$

Empirical Bootstrap

The empirical bootstrap method uses the **empirical CDF** as the estimate of F:

$$\hat{F}(a) = \frac{\sum_{i=1}^{n} \mathbb{1}_{X_i \le a}}{n}$$

where $\mathbb{1}_A$ is the indicator variable (Bernoulli) where

$$\mathbb{1}_A = \begin{cases} 1, & \text{if event } A \text{ occurs} \\ 0, & \text{if event } A \text{ does not occur} \end{cases}$$

Sampling $X_1, X_2, ..., X_n$ from F is replaced with sampling $X_1^*, X_2^2, ..., X_n^*$ from $\hat{F}$. This is achieved by taking a **random sample of size n, with replacement, from the original data set $x_1, x_2, ..., x_n$.**

Consider the following data set:

0.56	0.66	1.42	0.93
6.74	0.02	0.21	3.05

This data set has mean $\bar{x}_8 = 1.7$

Question 1:

Approximate the sampling distribution of $\bar{X}_8$ using empirical bootstrap method.

Perform one iteration of empirical bootstrap first:

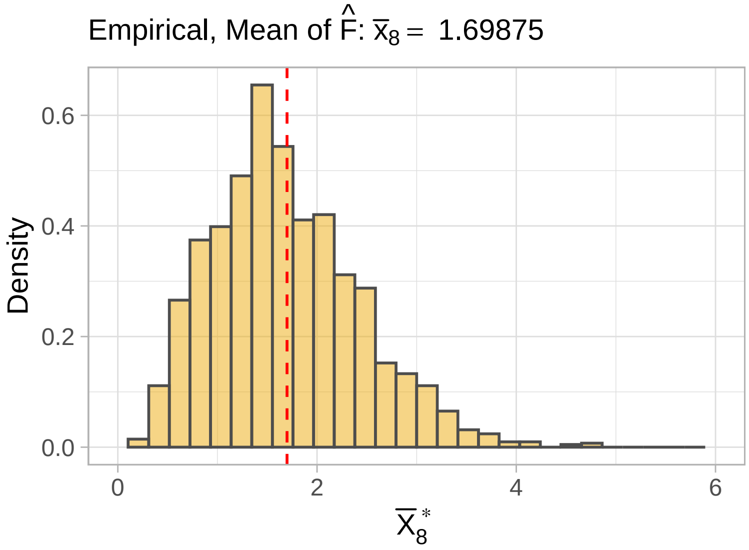
```r
set.seed(238)

#Mini Code: Sample with replacement:

### FILL THIS IN ###
boot.dat <- sample(og.data, 8,T) #store empirical bootstrap samples, og.data
boot.mean <- mean(boot.dat) #point estimate of x-bar for first bootstrap sample
```

Our first bootstrapped sample mean is 2.92. We repeat this a large number of times B, (the number of bootstrap samples) then store each bootstrap point estimate to construct the bootstrapped distribution of $\bar{X}_8$. We can do this in two ways:

- Writing a for-loop to run this iteratively until you have B resamples
- Store the data into a matrix and use the apply function. This can be advantageous for troubleshooting and verifying your results with the bootstrapped data since all bootstrap samples are saved!

```r
set.seed(20250224) # Ensures our results are reproducible

#For-Loop Method:
B = 2000 # num of bootstrap samples to produce
n = length(og.data)
emp.means <- c() # empty vector that will store each bootstrapped mean

for (i in 1:B){ #run bootstrap iteratively until you have B amount or you can store in matrix
  ### FILL THIS IN ##
  #perform empirical sample
  emp.sample <- sample(og.data, n, T)
  # calcualte sample mean and store in vec emp.means at ith position
  emp.means[i] <- mean(emp.sample)
}

#Matrix Method:
# create matrix w all bootstrap samples
# use apply function to get the means of each sample (each row)
  ### FILL THIS IN ##
all_samples <- matrix(sample(og.data, B*n, T), ncol=8) # each row has ith bootstrap sample
emp.mean2 <- apply(all_samples, 1, mean) # margin of 1 is rows, 2 is columns
```

Empirical Bootstrap Plot



I use colour codes in `ggplot` from this website: http://sape.inf.usi.ch/quick-reference/ggplot2/colour

---

It turns out that the data provided was a sample drawn from the $Exp(\theta)$ distribution. Can we produce improved bootstrap samples with this information?

Parametric Bootstrap

The parametric bootstrap method makes use of a partially specified model and a parameter estimate $\hat{\theta}$ to estimate the model distribution. i.e. $\hat{F} = F_{\hat{\theta}}$ is used to estimate F_θ

Sampling $X_1, X_2, ..., X_n$ from F_θ is replaced with sampling $X_1^*, X_2^2, ..., X_n^*$ from $F_{\hat{\theta}}$.

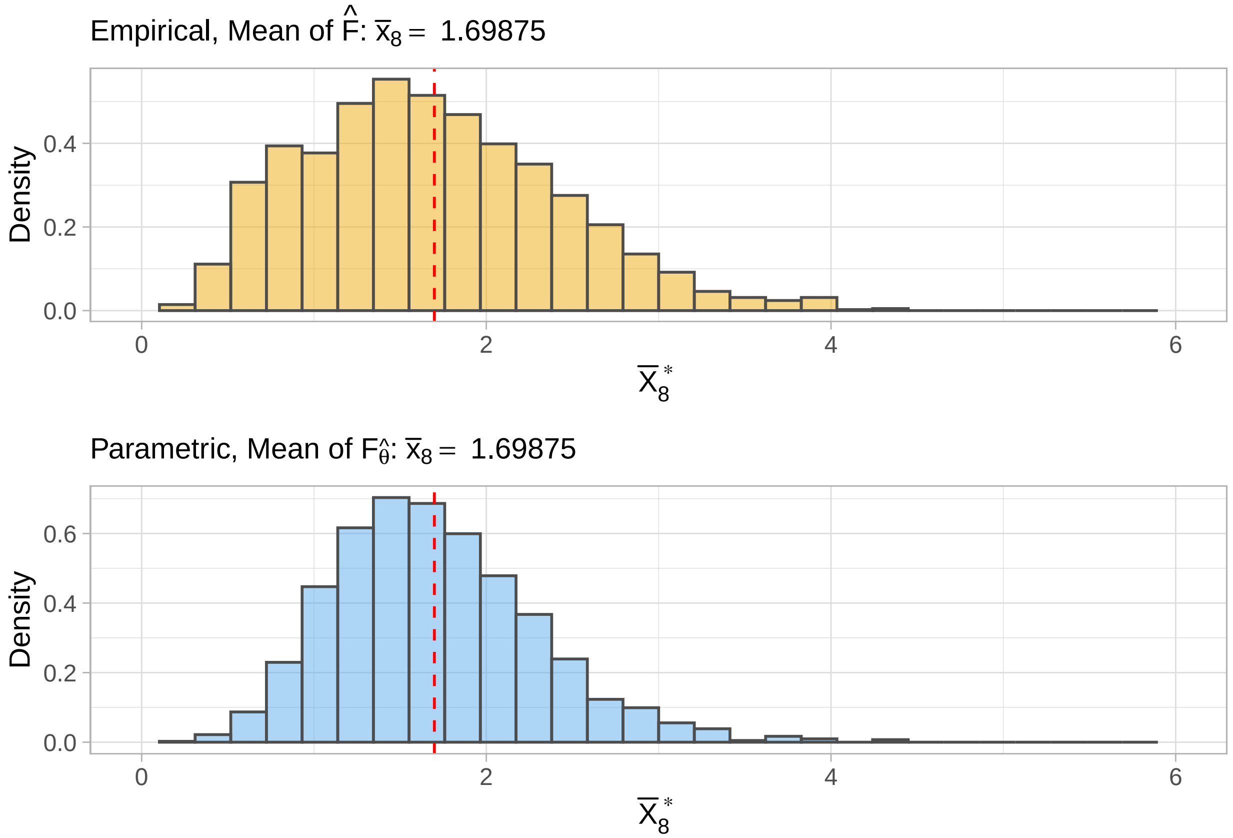
Question 2

Produce and compare plots of bootstrap distribution of the estimator from empirical and parametric bootstrap procedures. What do you notice about the shapes of the distributions?

Note: Empirical and parametric bootstrap differ only in *how we estimate* the model distribution F.

```
### Let's repeat empirical bootstrap alongside parametric bootstrap
# we believe data comes from exp dist. w theta unknown
# Parametric Bootstrap:
# Estimated model parameter
theta.hat <- mean(og.data) #estimate F with Exp(theta.hat)
#can show the mom and ml estimator for theta in exp(theta) is x-bar

para.means <- numeric() # store parametric sample means
emp.means <- numeric() # store empiricial sample means

for (i in 1:B){
 ### FILL THIS IN ##
  #empirical boot again
  emp.sample <- sample(og.data, n, T)
  emp.means[i] <- mean(emp.sample)

  #parametric boot, the difference is that we sample from exp.
  para.sample <- rexp(n, rate = 1/theta.hat) #rate is lambda default, which is why have to rename
  para.means[i] <- mean(para.sample)

}
```

Plot: Parametric Bootstrap vs Empirical Bootstrap


Empirical, Mean of $\hat{F}$: $\bar{x}_8 =$ 1.69875

Parametric, Mean of $F_{\hat{\theta}}$: $\bar{x}_8 =$ 1.69875

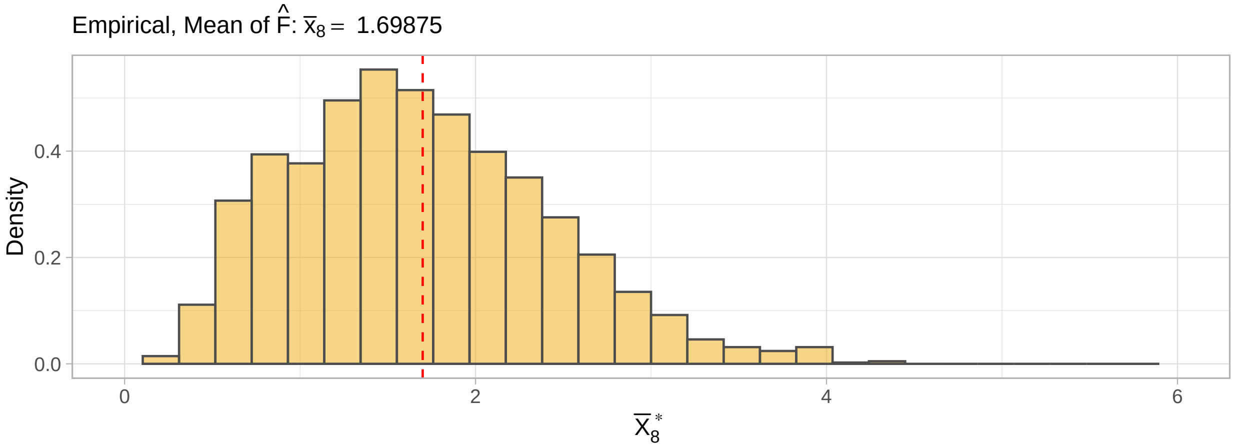
COMPARISON: Empirical bootstrap is based on only 8 data points, but we see that between empirical and parametric distributions the shapes are quite similar. Empirical is not too much worse. They have same centre (because using same sample mean). Both are right skewed, similar ranges in values of $\bar{X}_8$. The shape is smoother in parametric because we could boostrap from the distribution directly, while empirical always sampling from our dataset, so less diverse values. Empirical bootstrap is larger in variance because bigger distribution.

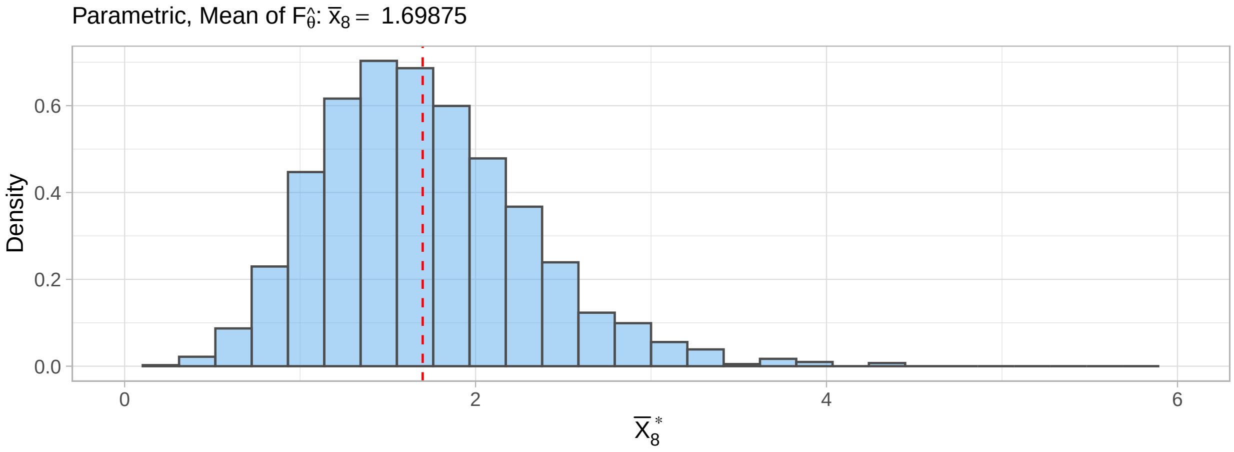
Question 3

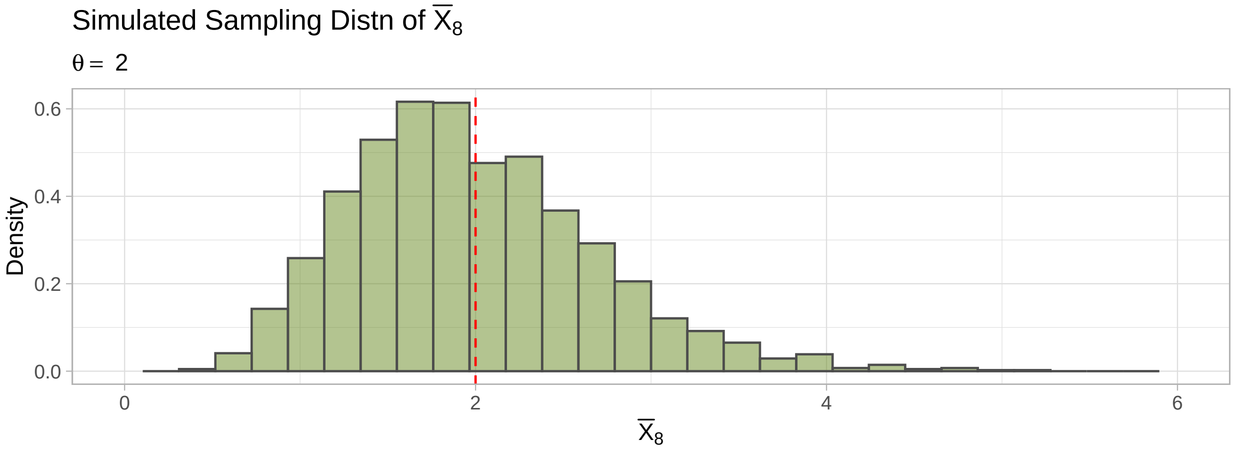
Suppose the data was actually sampled from an $Exp(\theta = 2)$ distribution. How do the bootstrapped distributions compare with one constructed via simulation? Empirical and parametric perform pretty well together. But compared to actual simulation $\bar{X}$?

```
# True model parameter
theta=2
#sampling from the pop dist.
#In simulation:
# 1. We draw NEW data from the population distribution
# 2. Compute the sample means, repeat B times
# 3. Plot the sample means as an approximation to the sampling distribution
# of X-bar_8
B = 2000
sim.means <- numeric()
for (i in 1:B){
 ### FILL THIS IN ##
  sim.sample <- rexp(n, rate = 1/2)
  sim.means[i] <- mean(sim.sample)
}
```

Plot of Bootstraps vs Simulation


Empirical, Mean of $\hat{F}$: $\overline{x}_8 =$ 1.69875


Parametric, Mean of $F_{\hat{\theta}}$: $\overline{x}_8 =$ 1.69875


Simulated Sampling Distn of $\overline{X}_8$

$\theta =$ 2

Notable Differences:

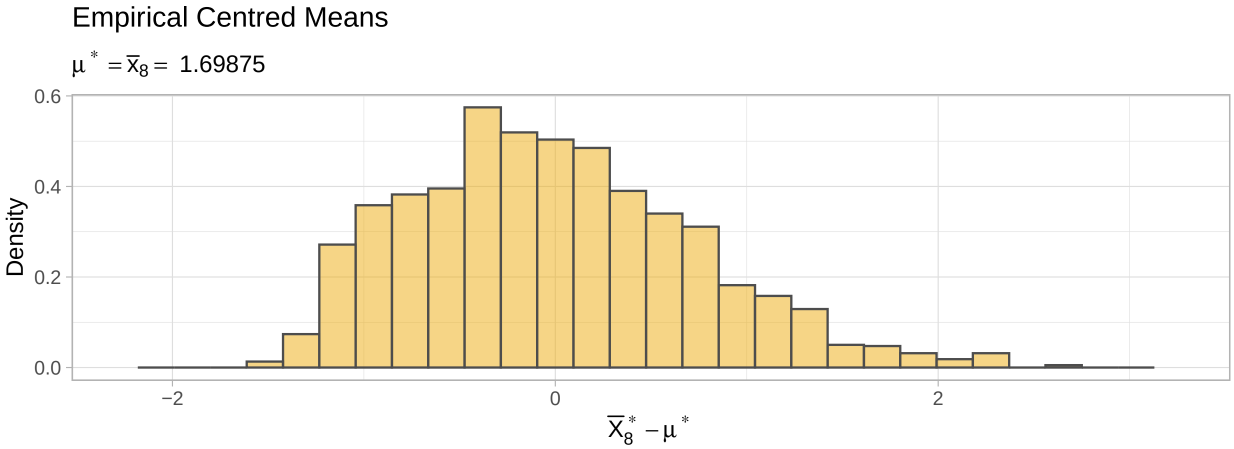
- In empirical and parametric bootstrap, the distribution for the sample mean is centred on the *sample mean of the original data set* since both $\hat{F}$ and $F_{\hat{\theta}}$ are centred on $\bar{x}_8$
- The simulated distribution of the sample mean is centred on the true parameter $\theta = 2$
- Bootstrap sample mean is to the left of the true mean.
- Bootstrap samples are going to be biased toward the estimated means. Because of this difference, it's better to examine bootstrapped centred means instead of bootstrapped means.
- Rather than focus on the values of the estimator (which is influenced by our data), it is more productive to examine how variable bootstrap estimates are from the "mean" of $\hat{F}$
- In other words, we want information about $\bar{X}_8 - \mu$ instead of $\bar{X}_8$. We use $\bar{X}_8^* - \mu^*$ to approximate behaviour of $\bar{X}_8 - \mu$.
- parametric is sensitive to the point estimate of the parameter because of bias. For bias, try to subtract bias from centre, for variance issues, standarize w z-score
- **Note:** μ^* is used to denote the expected value of the bootstrap distribution. In this case,
    - $\mu^* = \bar{x}_8$ in empirical bootstrap, and
    - $\mu^* = \hat{\theta} = \bar{x}_8$ in parametric bootstrap.

Centred Means

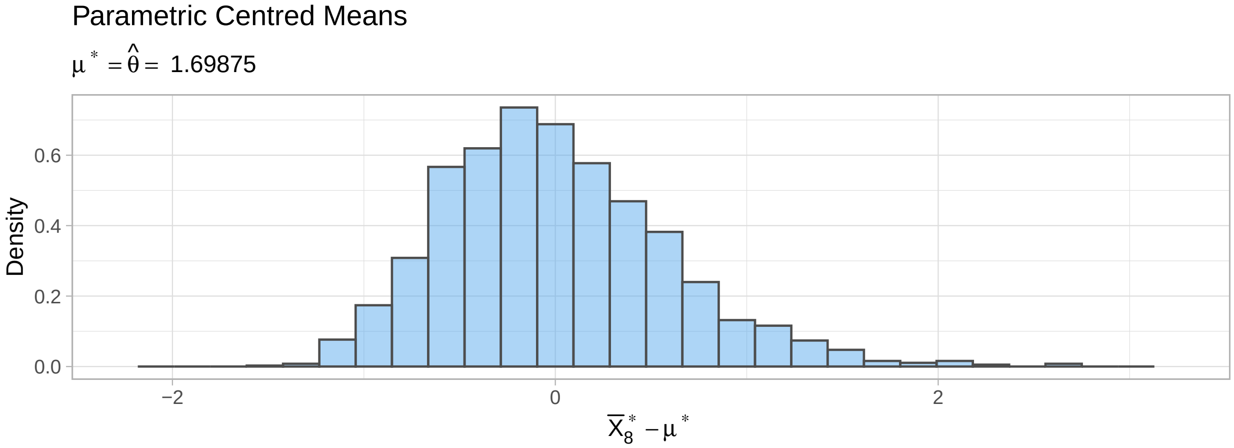
Question 4

Produce a graph of bootstrapped centred means and compare them to those produced in simulation. What do you notice about the plots of deviations $\bar{X}_8 - \mu^*$?

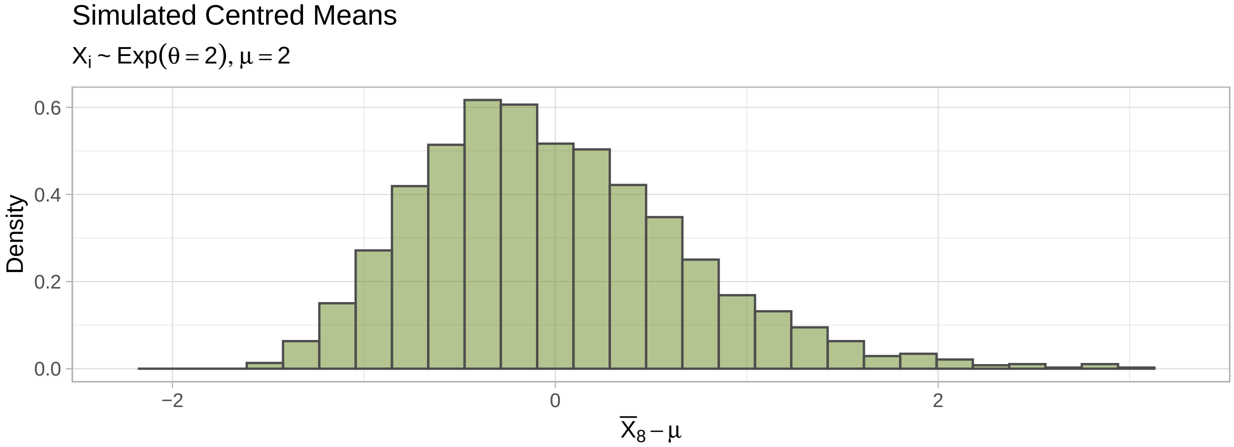
```
# Centred Means
# take all of the simulated means and convert to distances from the centre of their respective F-hat or
centred.values <- tibble(emp.c = emp.means - mean(og.data),
                         para.c = para.means - mean(og.data),
                         sim.c = sim.means - theta)
```


Empirical Centred Means

$\mu^* = \bar{x}_8 = 1.69875$


Parametric Centred Means

$\mu^* = \hat{\theta} = 1.69875$


Simulated Centred Means

$X_i \sim \text{Exp}(\theta = 2), \mu = 2$

COMPARISON: Para means have a dist. most similar in shape to the simulated values, but a smaller variance because variance is θ^2 which is 1.7^2 in parametric model while 2^2 in simulation. Empirical centred vals are more conservative, and has a wider variance. All three are pretty similar.

Question 5

Based on the bootstrap distribution, estimate the probability that the estimator yields a value within 0.5 units of the parameter.

```r
# Estimate: Probability of estimating within 0.5 units of parameter
#how likely will we see a random sample mean that is within the true param.
emp.prob <- sum(abs(centred.values$emp.c) <= 0.5)/B # relative frequency
para.prob <- sum(abs(centred.values$para.c) <= 0.5)/B
sim.prob <- sum(abs(centred.values$sim.c) <= 0.5)/B
```

(i) By empirical bootstrapping, we estimate this probability to be 48.8%. We should go with this one, the minimum.
(ii) By parametric bootstrapping, we estimate this probability to be 61.3%.
(iii) By simulation (which should be most accurate to true probability), we estimate this probability to be 52.5%.

Exercises

1. Consider the following data on average daily temperatures in February 2025 in Toronto provided by the Government of Canada, measured in degrees Celsius:

-11.3	-7.8	-1.3	-5.5	-7.4	-2.1	-5.4
-5.7	-3.1	-7.1	-7.3	-7.0	-5.8	-6.6
-4.4	-7.3	-10.3	-10.5	-9.4	-9.5	-9.3

You are tasked with investigating the variability in the daily temperatures in the month of February for the city. In particular, if the temperature fluctuation is high, the city may consider notifying residents to keep their taps running with a trickle of warm water to prevent sudden expansions/contractions of piping which could lead to leaky pipes. The data for the first 21 days of February are stored in `mean.temp`.

a) Find the point estimate for the variance in daily temperatures.

b) How many bootstrapped data sets can be generated from this data?

c) Construct a bootstrapped distribution for sample variance, using $B = 2000$. What is the range of sample variances that occur 90% of the time? (Use R to find the appropriate percentiles to answer this question! What percentiles should you use?)

d) This might not be useful as bootstrapping produces sampling distributions that are centred on our data. Instead, examine the sampling distribution of normalized variances: $\frac{S_n^2}{\sigma^2}$. How do you mimic S_n^2 in bootstrapping? How about σ^2?

e) Produce the bootstrapped distribution of normalized variances, using $B = 2000$. Plot the density histogram with clear labels and titles.

f) Find the interval of ratios that occur 90% of the time (Use R to find the appropriate percentiles to answer this question! What percentiles should you use?). How can you interpret this interval to infer how well the sample variance measures against the true variance that you are trying to estimate?

```
set.seed(238)
feb_sample = c(-11.3,   -7.8,   -1.3,   -5.5,   -7.4,   -2.1,   -5.4,
-5.7,   -3.1,   -7.1,   -7.3,   -7.0,   -5.8,   -6.6,
-4.4,   -7.3,   -10.3, -10.5,   -9.4,   -9.5,   -9.3)
# question a:
sample_var <- var(feb_sample)

# question b:
n <- length(feb_sample)
n^n # all possible data sets
```

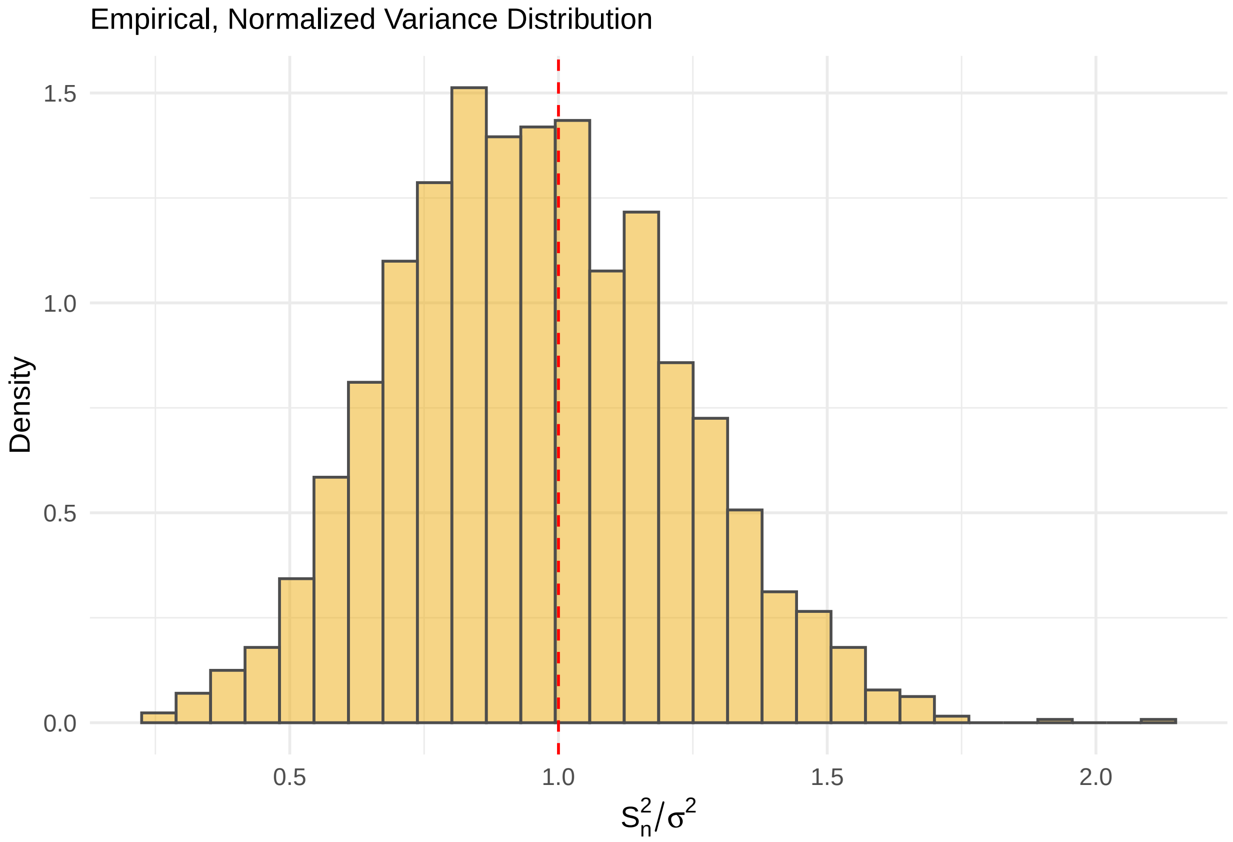
```
## [1] 5.842587e+27
```

```
choose(2*n - 1, n) # useful data sets
```

```
## [1] 269128937220
```

```
# question c:
B <- 2000
boot_variances <- c()
for (i in 1:B) {
  boot_sample <- sample(feb_sample, n, replace = TRUE)
  boot_variances[i] <- var(boot_sample)
}


# Question is asking for the middle 90%, which means you need to leave out the tails.
relative_freq_90 <- quantile(boot_variances, c(0.05, 0.95))
# should be an inclusive range [q_05, q_95]

# question d:
# sigma^2 is the variance of the original sample, which we calculated already, and S^2_n are the bootst
ratio_bootstrap <- c()
for (i in 1:B) {
  ratio_bootstrap[i] <- (boot_variances[i]/sample_var)
}

# question e:

library(ggplot2)
library(latex2exp)  # For TeX labels

# Create the plot
norm_var_plot <- ggplot(data.frame(ratio_bootstrap), aes(x = ratio_bootstrap)) +
  geom_histogram(aes(y = after_stat(density)), bins = 30,
                 fill = 'darkgoldenrod2', color = 'grey30', alpha = 0.5) +
  geom_vline(aes(xintercept = 1), linetype = "dashed", color = 'red') +
  labs(x = TeX(r'($S_n^2 / \sigma^2$)'),
       y = 'Density',
       subtitle = TeX(paste("Empirical, Normalized Variance Distribution"))) +
  theme_minimal()
```

```r
# Display the plot
print(norm_var_plot)
```



```r
# question f:
 relative_freq_norm_90 <- quantile(ratio_bootstrap, c(0.05, 0.95))
```

2. We return to the system interfailure times (measured in CPU seconds) data set from Week 2. We previously established that an $Exp(\theta)$ would be a suitable model distribution for the data. In this problem, you will investigate the **median failure time** instead of the mean failure time.

30	113	81	115	9	2	91	112	15	138	50	77	24	108	88
670	120	26	114	325	55	242	68	422	180	10	1146	600	15	36
4	0	8	227	65	176	58	457	300	97	263	452	255	197	193
6	79	816	1351	148	21	233	134	357	193	236	31	369	748	0
232	300	365	1222	543	10	16	529	379	44	129	810	290	300	529
281	160	828	1011	445	296	1755	1064	1783	860	983	707	33	868	724
2323	2930	1461	843	12	261	1800	865	1435	30	143	108	0	3110	1247
943	700	875	245	729	1896	447	386	446	122	990	948	1082	22	75
482	5509	100	10	1071	371	790	6150	3321	1045	648	5485	1160	1864	4116

a) Which of empirical or parametric bootstrap do you think would be the more appropriate procedure in this context? Explain why.

b) Why is a simulation-based approach not possible here?

c) Perform empirical bootstrap on the data set here and plot the bootstrap distribution of **centred medians**. Use ‘B = 1000‘. *Note: Since median is a location measure, centering should involve a difference of measures*.

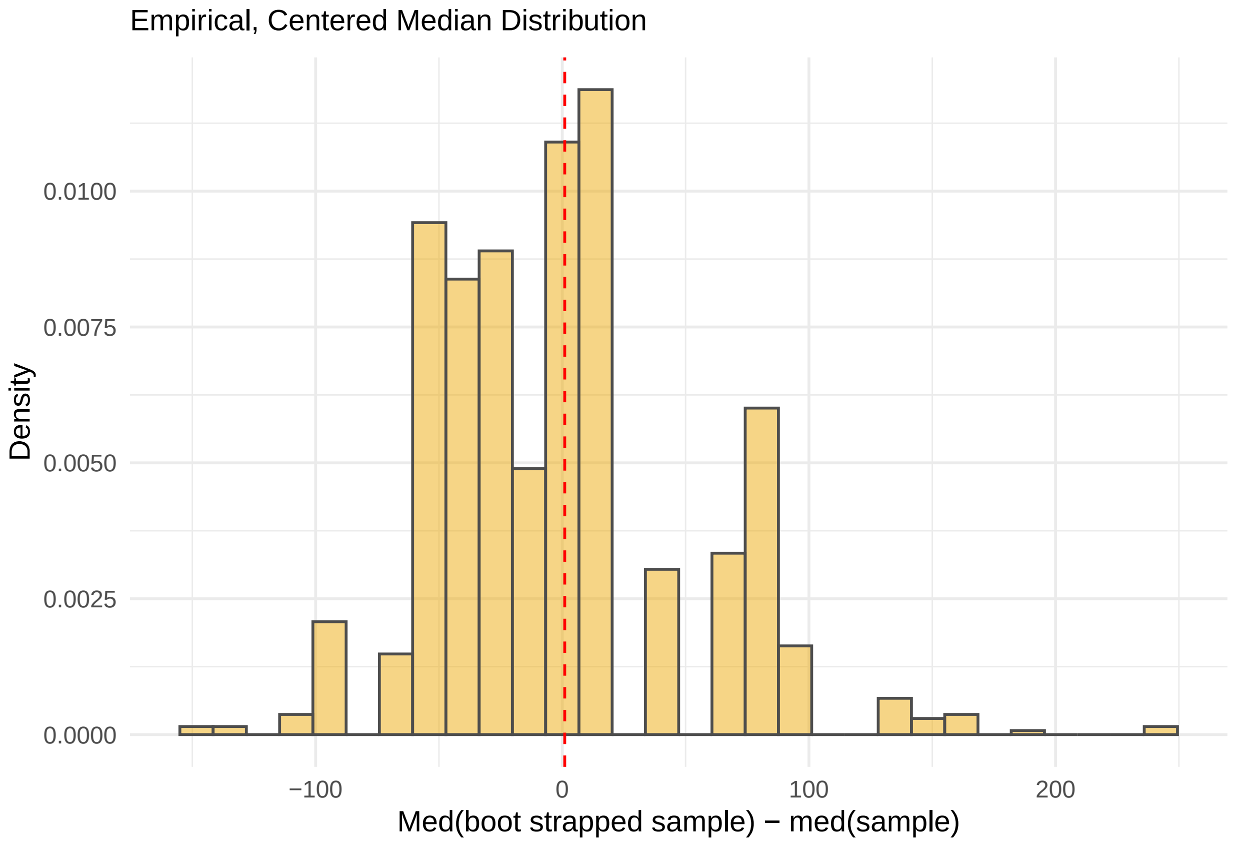
d) In order to perform parametric bootstrap, you need the following:

- The parameter estimate $\hat{\theta}$
- The estimated measure of interest

Find both of these estimates.

e) Perform parametric bootstrap and plot the distribution of **centred medians**. Use ‘B = 1000‘. How does the parametric bootstrap compare to the empirical bootstrap, visually?

```r
set.seed(238)
# question a:
# Parametric bootstrap is more appropriate because we are given a partially specified model with theta 

# question b:
# The problem here is that we do not know the true parameter, rather, only an estimate (theta hat), whi

# question c:
sample_failures <- c(30, 113, 81, 115, 9, 2, 91, 112, 15, 138,
              50, 77, 24, 108, 88, 670, 120, 26, 114, 325,
              55, 242, 68, 422, 180, 10, 1146, 600, 15, 36,
              4, 0, 8, 227, 65, 176, 58, 457, 300, 97,
              263, 452, 255, 197, 193, 6, 79, 816, 1351, 148,
              21, 233, 134, 357, 193, 236, 31, 369, 748, 0,
              232, 300, 365, 1222, 543, 10, 16, 529, 379, 44,
              129, 810, 290, 300, 529, 281, 160, 828, 1011, 445,
              296, 1755, 1064, 1783, 860, 983, 707, 33, 868, 724,
              2323, 2930, 1461, 843, 12, 261, 1800, 865, 1435, 30,
              143, 108, 0, 3110, 1247, 943, 700, 875, 245, 729,
              1896, 447, 386, 446, 122, 990, 948, 1082, 22, 75,
              482, 5509, 100, 10, 1071, 371, 790, 6150, 3321, 1045,
              648, 5485, 1160, 1864, 4116)
B <- 1000
n <- length(sample_failures)
sample_median <- median(sample_failures)
bootstrap_medians_centre <- c()
for (i in 1:B) {
  bootstrap_medians_centre[i] <- median(sample(sample_failures, n, replace = TRUE) - sample_median)
}
centred_medians_plot <- ggplot(data.frame(bootstrap_medians_centre), aes(x = bootstrap_medians_centre))
  geom_histogram(aes(y = after_stat(density)), bins = 30,
                 fill = 'darkgoldenrod2', color = 'grey30', alpha = 0.5) +
  geom_vline(aes(xintercept = 1), linetype = "dashed", color = 'red') +
  labs(x = TeX(r'(Med(boot strapped sample) - med(sample))'),
       y = 'Density',
       subtitle = TeX(paste("Empirical, Centered Median Distribution"))) +
  theme_minimal()

# Display the plot
print(centred_medians_plot)
```



```
# question d
# too tired to type this out but find theta hat through MLE or MOM, call it theta_hat
# same bootstrap process except samples are coming from med(rexp(n, rate = 1/theta_hat))
```