MAST20005/MAST90058: Week 10 Lab

Goals: (i) Properties of order statistics; (ii) Confidence intervals for quantiles; (iii) An introduction to the bootstrap and an example of its use.

1 Order statistics

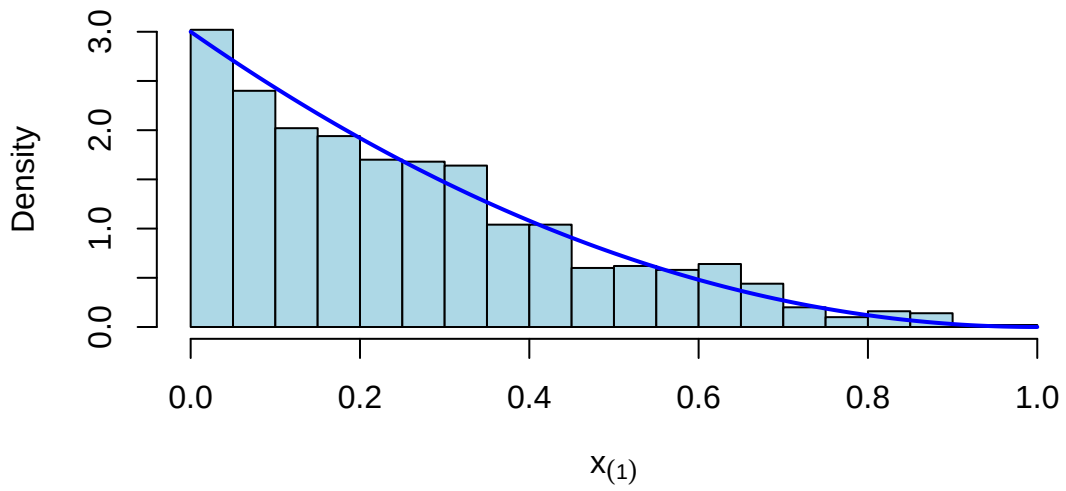
Let $X_{(1)} < X_{(2)} < X_{(3)}$ be order statistics of a random sample X_1, X_2, X_3 from the the uniform distribution $\text{Unif}(\theta - 0.5, \theta + 0.5)$. Three possible estimators for the median are the sample mean $W_1 = \bar{X}$, the sample median $W_2 = X_{(2)}$, and the midrange $W_3 = (X_{(1)} + X_{(3)})/2$.

1. The theoretical pdf for the smallest order statistic, $X_{(1)}$, is

$$g_1(x) = 3\left(1 - F(x)\right)^2 f(x) = 3\left(1 - x\right)^2, \quad \theta - 0.5 < x < \theta + 0.5.$$

Simulate the distribution of the first order statistic $X_{(1)}$ and compare the resulting density with the above theoretical density, using some particular value for θ (for example $\theta = 0.5$).

```
theta <- 0.5
x1.simulated <- numeric(1000)  # initialise an empty vector
for (i in 1:1000) {
    x <- runif(3, theta - 0.5, theta + 0.5)
    x1.simulated[i] <- min(x)
}

g <- function(x)
    3 * (1 - x)^2

hist(x1.simulated, breaks = 20, freq = FALSE, col = "lightblue",
     xlim = c(0, 1), ylim = c(0, g(0)),
     main = NULL, xlab = expression(x[(1)]))
curve(g, from = 0, to = 1, add = TRUE, col = "blue", lwd = 2)
```



2. Simulate 100 samples of size 3 from the uniform model above and calculate and store values of W_1, W_2 and W_3.

```
w.simulated <- matrix(nrow = 100, ncol = 3)  # initialise an empty matrix
for (i in 1:100) {
    x <- runif(3, theta - 0.5, theta + 0.5)
    x <- sort(x)
    w1 <- mean(x)
    w2 <- x[2]
    w3 <- (x[1] + x[3]) / 2
    w  <- c(w1, w2, w3)
    w.simulated[i, ] <- w
}
```

3. Compare the values of the sample means and sample variances for W_1, W_2 and W_3. Which of these statistics is the best estimator of θ?

```
# Compute `mean` and `var` for each column of `w.simulated`.
means <- apply(w.simulated, 2, mean)
vars  <- apply(w.simulated, 2, var )
means
```

```
## [1] 0.4994224 0.4843060 0.5069806
```

```
vars
```

```
## [1] 0.02755129 0.05641642 0.02368700
```

Here, `apply()` runs the given function (3rd argument) to each column of the given matrix (1st argument). If the second argument has value `1`, then it would run the given function to each row instead.

4. Simulated estimates are affected by error due to random sampling, thus one could question whether 100 runs are enough to give reliable results. Therefore, it is typical to either run a very large number of simluations or otherwise present simulation results in terms of interval estimates. By the Central Limit Theorem, the means from the simulations follow a normal distribution. That allows us to calculate 95% confidence intervals for $E(W_1)$, $E(W_2)$ and $E(W_3)$ in the following way:

```
# CI for E(W1).
means[1] + c(-1, 1) * qnorm(0.975) * sqrt(vars[1]) / sqrt(100)
```

```
## [1] 0.4668898 0.5319550
```

Note: It turns out that $E(W_1) = E(W_2) = E(W_3) = 0.5$ (you may check this as a homework problem). Clearly, the intervals above contain the true value 0.5.

2 Confidence intervals for quantiles

Let $X \sim \text{Unif}(0,1)$ and consider a random sample of size 11 from X. In the lectures we saw that if m is the median and $X_{(1)}, \ldots, X_{(n)}$ are the order statistics then

$$\Pr(X_{(i)} < m < X_{(j)}) = \sum_{k=i}^{j-1} \binom{n}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{n-k}.$$

We will check this formula using R by computing confidence intervals for the median of X.

1. Use `qbinom()` to compute quantiles of the Bi(11, 0.5) distribution (e.g. find $\pi_{0.975}$ so that $\Pr(X \leqslant \pi_{0.975}) \approx 0.975$).

```
qbinom(c(0.025, 0.975), 11, 0.5)
```

```
## [1] 2 9
```

   Note that these are approximations as the distribution is discrete. However, they can be used as endpoints of the desired confidence interval.

2. Determine $\Pr(X_{(2)} < m < X_{(9)})$

```
pbinom(8, 11, 0.5) - pbinom(1, 11, 0.5)
```

```
## [1] 0.9614258
```

   Note that the confidence interval is only approximate. However, it is still useful since it ensures a confidence level slightly larger than 95%.

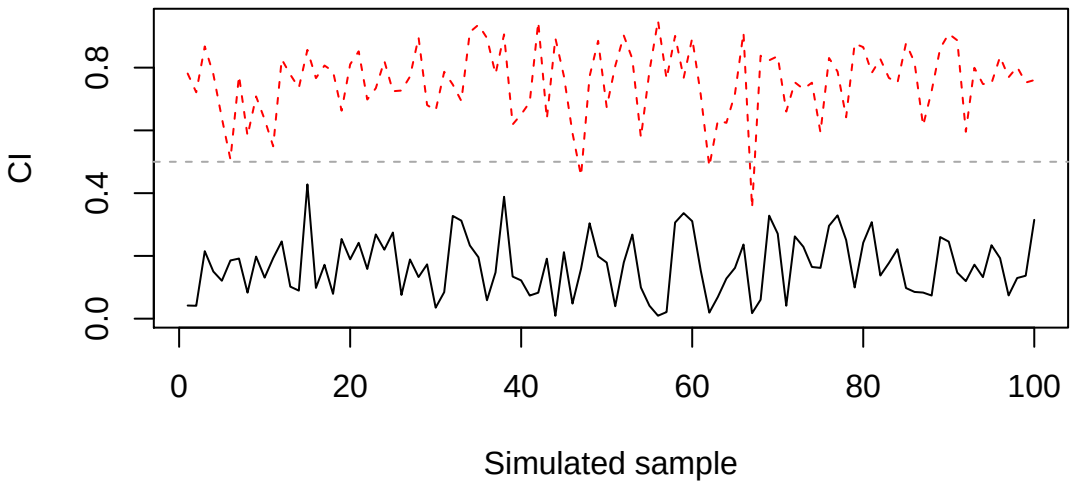
3. The R command `X <- runif(11)` simulates 11 observations from X, while `sort()` computes the order statistics. We automate calculation of $X_{(2)}$ and $X_{(9)}$ by the following function:

```
f <- function() {
    X <- runif(11)
    Y <- sort(X)
    c(Y[2], Y[9])
}
f()  # try it out
```

```
## [1] 0.1012517 0.9010142
```

4. Next we check the coverage probability for the interval above by simulation.

```
nsimulations <- 100
C <- t(replicate(nsimulations, f()))
matplot(C, type = "l", xlab = "Simulated sample", ylab = "CI")
abline(c(0.5, 0), lty = 2, col = "darkgrey")
```




```
mean((C[, 1] < 0.5) & (0.5 < C[, 2]))
```

```
## [1] 0.97
```

The above code computes the proportion of simulated intervals that contain the true median value 0.5. Is this close to your answer from question 2, above? The associated plot shows the confidence interval endpoints for each simulation run. If you want more precision, repeat with `nsimulations = 1000`.

3 Bootstrap*

3.1 Introduction

The bootstrap is a computational technique to approximate the sampling distribution of almost any statistic. It is an example of a *resampling method*, which refers to the fact that it involves taking samples from the original sample (and so multiple times). The resulting approximate distribution can be used to obtain confidence intervals or for hypothesis testing.

Suppose we have an iid sample $X_1, \ldots, X_n$ from some unknown distribution. Our main interest is to find the distribution of some statistic, say $\hat{\theta}$ (e.g. sample median, sample variance, estimate of a regression coefficient). Such a distribution can be approximated by the following simple steps:

1. Obtain a new sample, $X_1^*, \ldots, X_n^*$, by drawing **with replacement** from the original observations $X_1, \ldots, X_n$.

2. Using the sample $X_1^*, \ldots, X_n^*$, compute and store the statistic $\hat{\theta}^*$.

3. Repeat steps 1 and 2 many times, say B, where B might be 1 000, 2 000 or even 10 000, thus obtaining $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$. These are known as the *bootstrapped statistics*.

4. The empirical distribution of the bootstrapped statistics can be regarded as an approximation of the distribution for the statistic $\hat{\theta}$ computed from the original sample (thus, they can be used to find confidence intervals, etc.).

Advanced theory shows that the bootstrap approximation works well even if n is quite small. This is in contrast to procedures relying on the Central Limit Theorem, which typically require much larger samples. For example, let us consider iid samples from the Cauchy distribution with pdf,

$$f(x \mid \theta) = \frac{1}{\pi[1 + (x - \theta)^2]}, \quad -\infty < x < \infty.$$

The parameter θ represents the median. In R we can generate $n = 25$ samples using $\theta = 5$ as follows:

```
x <- rcauchy(25, location = 5)
x

##  [1]   5.3848741   5.5720126  -1.6699114 -13.2179377   4.3494096
##  [6]   5.9071336  82.0045986   5.6815994  -0.4244848   6.2010912
## [11]  -1.0851728   4.3676934   4.5136773   4.0890473  10.7211124
## [16]  -2.3434305  -4.1760812   6.1778866   9.5474368   0.2215683
## [21]   4.4807721   2.6996726   4.1524081   4.1804312   6.1526273
```

Interestingly, many values are between 3 and 7 so they resemble a sample from a normal distribution with mean 5 and standard deviation 1. However, note there there are many outliers represented by quite extreme values. These occur because the Cauchy distribution has very long tails. The presence of outliers suggests that $\bar{X}$ is not a very good estimator of the location.

Next compare the distribution of the sample mean $\bar{X}$ with that of the trimmed mean $\bar{X}_{\text{tr}}$ (mean without the most extreme observations in each tail). The statistics computed from the original sample are:

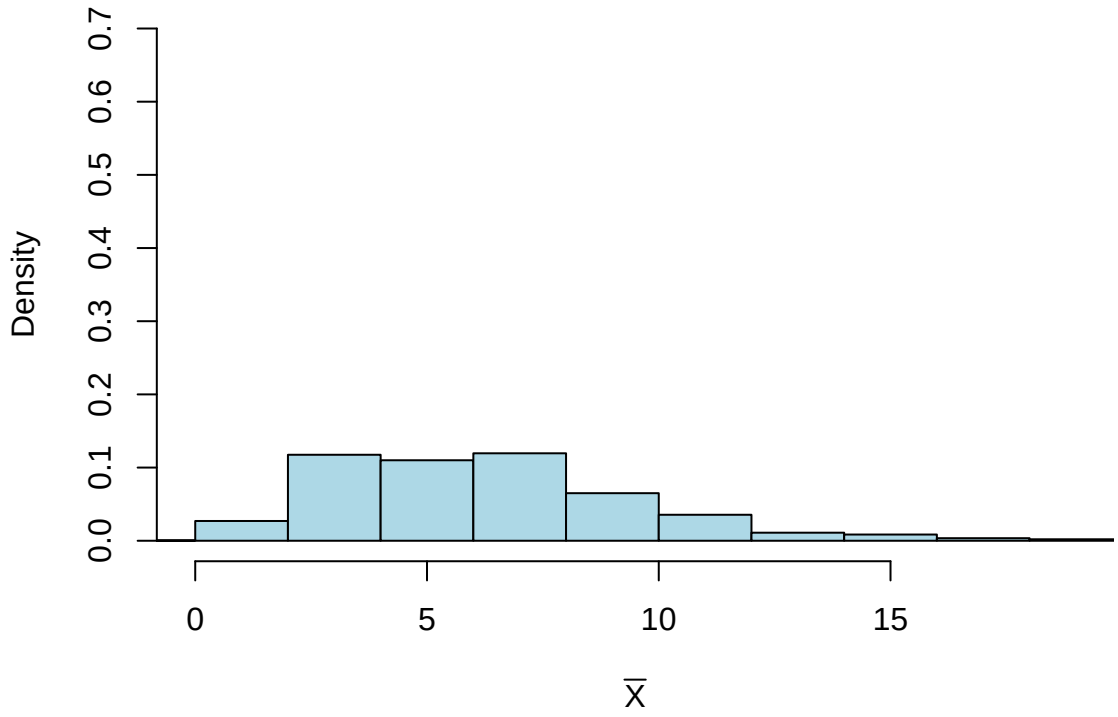
```
x.bar <- mean(x)
x.bar

## [1] 6.139521

x.bar.tr <- mean(x, trim = 0.35)  # exclude 35% of observations from each tail
x.bar.tr

## [1] 4.565592
```

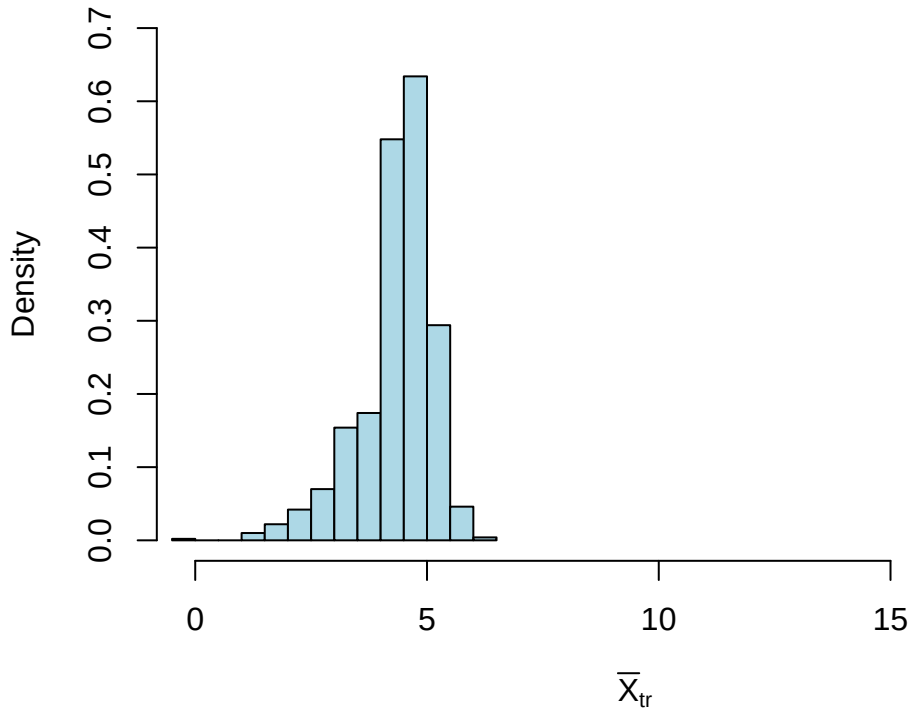
1. The statistics can be bootstrapped as follows:

```
B <- 1000
x.bar.boot    <- numeric(B)
x.bar.tr.boot <- numeric(B)
for (i in 1:B) {
    x.ast <- sample(x, size = 25, replace = TRUE)
    x.bar.boot[i]    <- mean(x.ast)
    x.bar.tr.boot[i] <- mean(x.ast, trim = 0.35)
}
```

2. Plot the distribution of the bootstrapped statistics, using a common scale:

```
xlim <- range(x.bar.boot, x.bar.tr.boot)
ylim <- c(0, 0.7)
par(mfrow = c(2, 1), mar = c(5.1, 4.1, 1, 1))
hist(x.bar.boot, xlab = expression(bar(X)), freq = FALSE,
     xlim = xlim, ylim = ylim, col = "lightblue", main = NULL)
hist(x.bar.tr.boot, xlab = expression(bar(X)[tr]), freq = FALSE,
     xlim = xlim, ylim = ylim, col = "lightblue", main = NULL)
```





3. To find a 95% confidence interval for θ we will use the percentile bootstrap method. This involves simply calculating the sample quantiles of the bootstrapped statistics; for example, the 2.5% and 97.5% sample quantiles in order to get a 95% confidence interval. In R, sample quantiles are computed using `quantile()`.

```
quantile(x.bar.tr.boot, c(0.025, 0.975))
```

```
##     2.5%    97.5%
## 2.221659 5.485469
```

The percentile bootstrap is a distribution-free method, since we make no specific assumptions about the underlying distribution.

3.2 Old Faithful geyser data

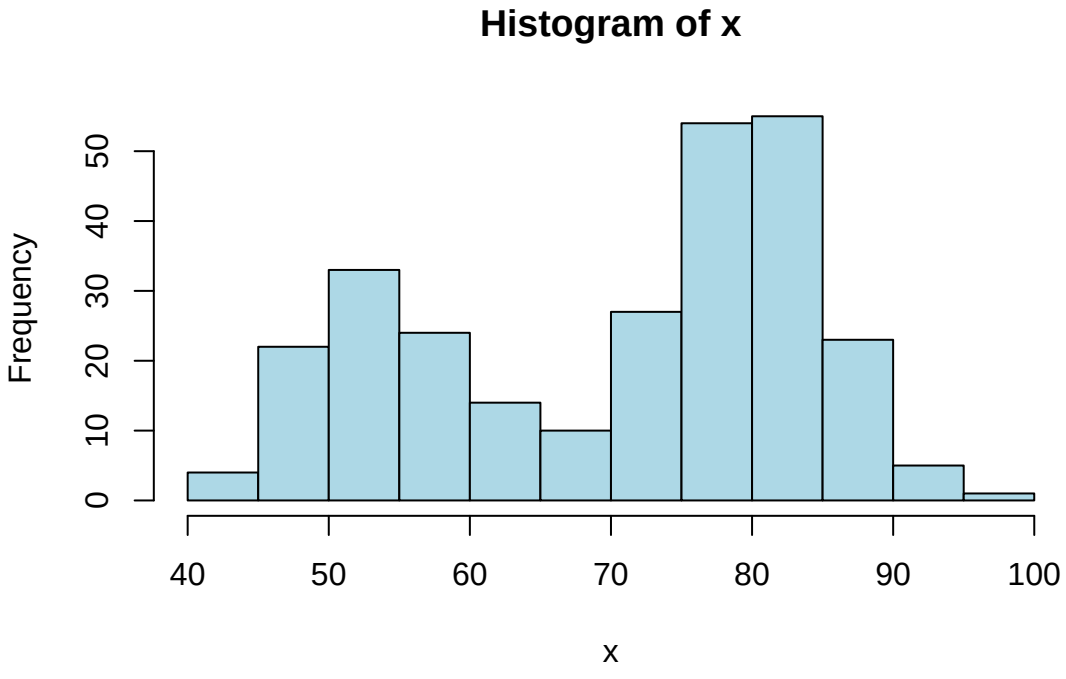
In this section we use the waiting times (in minutes) between the starts of successive eruptions of the Old Faithful geyser in Yellowstone National Park (Wyoming, USA). The data were collected continuously from 1 August until 15 August, 1985. We examine various features of this data using the bootstrap method. The data should be available in the data frame `faithful` in any standard installation of R.

1. The following computes the sample mean and shows a histogram of the waiting times.

```
x <- faithful$waiting
mean(x)
```

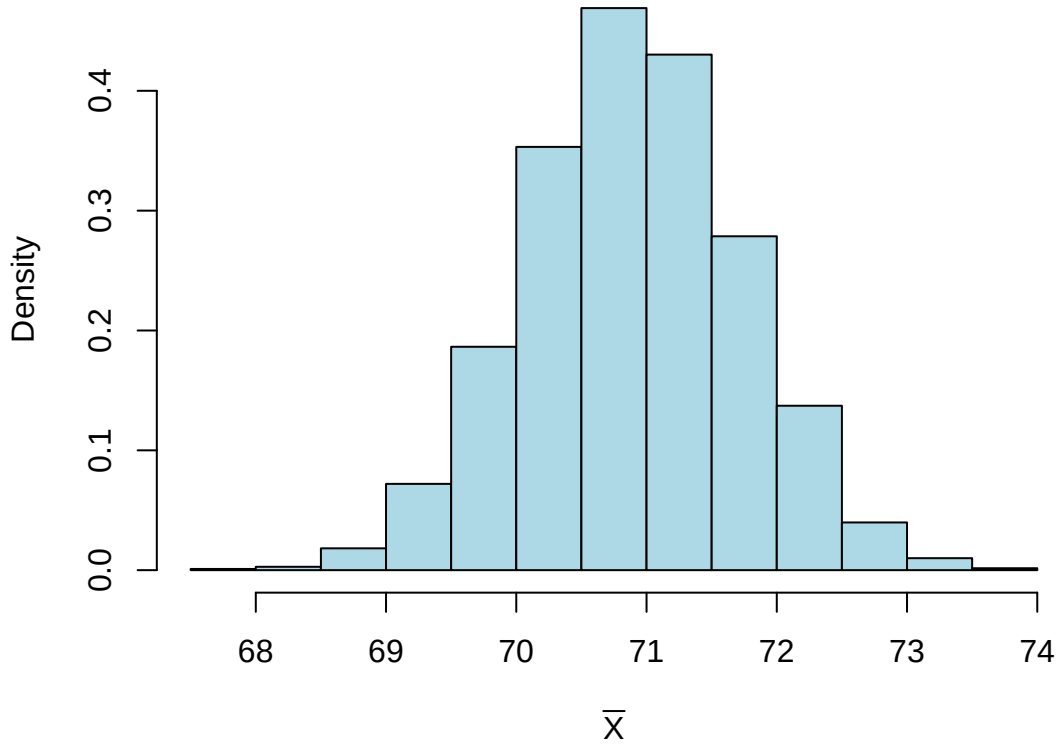
```
## [1] 70.89706
```

```
hist(x, col = "lightblue")
```



2. Generate 10 000 bootstrap replicates of $\bar{X}$ and plot the bootstrap distribution.

```
B <- 10000
x.bar.boot <- numeric(B)
for (i in 1:B) {
    x.ast <- sample(x, replace = TRUE)
    x.bar.boot[i] <- mean(x.ast)
}
hist(x.bar.boot, xlab = expression(bar(X)), freq = FALSE,
     col = "lightblue", main = NULL)
```



3. A 95% confidence interval for the true mean waiting time μ is obtained as follows

```
quantile(x.bar.boot, c(0.025, 0.975))

##     2.5%    97.5%
## 69.28676 72.51103
```

Note that this confidence interval uses no information about the true distribution generating the data. It is then easy to see how this procedure gets its name, because it is like "pulling yourself by your own bootstraps" with the empirical distribution acting as a bootstrap.

Exercises

1. Consider a random sample of size 4 from an exponential distribution with rate parameter 1. Simulate the distribution of the first order statistic. Draw a histogram of the simulated values and superimpose the theoretical pdf (which you will need to derive).

2. Consider the shifted exponential distribution with pdf:

$$f(x \mid \theta) = e^{-(x-\theta)} \quad (x > \theta).$$

   We discussed this in the lectures early on. Two estimators we proposed were $T_1 = \bar{X} - 1$ and $T_2 = X_{(1)} - \frac{1}{n}$. Using $\theta = 3$ and a sample size of $n = 10$, use simulations to show that both of these are unbiased and that T_2 has clearly smaller variance.

3. Consider the scenario in Section 1.

   (a) Consider the estimator $W_4 = X_{(3)} - 0.5$. Use simulations to show that it is biased.

   (b) Determine a value of a that makes $W_5 = X_{(3)} - a$ an unbiased estimator.

   (c) Use simulations to compare the variance of W_5 to that of W_1, W_2 and W_3.

4. Calculate a 95% confidence interval for the simulated coverage estimate in Section 2. Repeat for 1000 simulations.

5. Do question 5 from the tutorial problems. Also, find an approximate 95% confidence interval for the first quartile.

6. Consider the following random sample on X:

```
0.252, 0.287, 0.537, 0.511, 0.054,
0.022, 0.142, 0.021, 0.155, 0.241
```

   Calculate the statistic $T = 1/\bar{X}$. Suppose this is an estimator for some underlying parameter θ. Calculate a 95% confidence interval for θ using the percentile bootstrap.