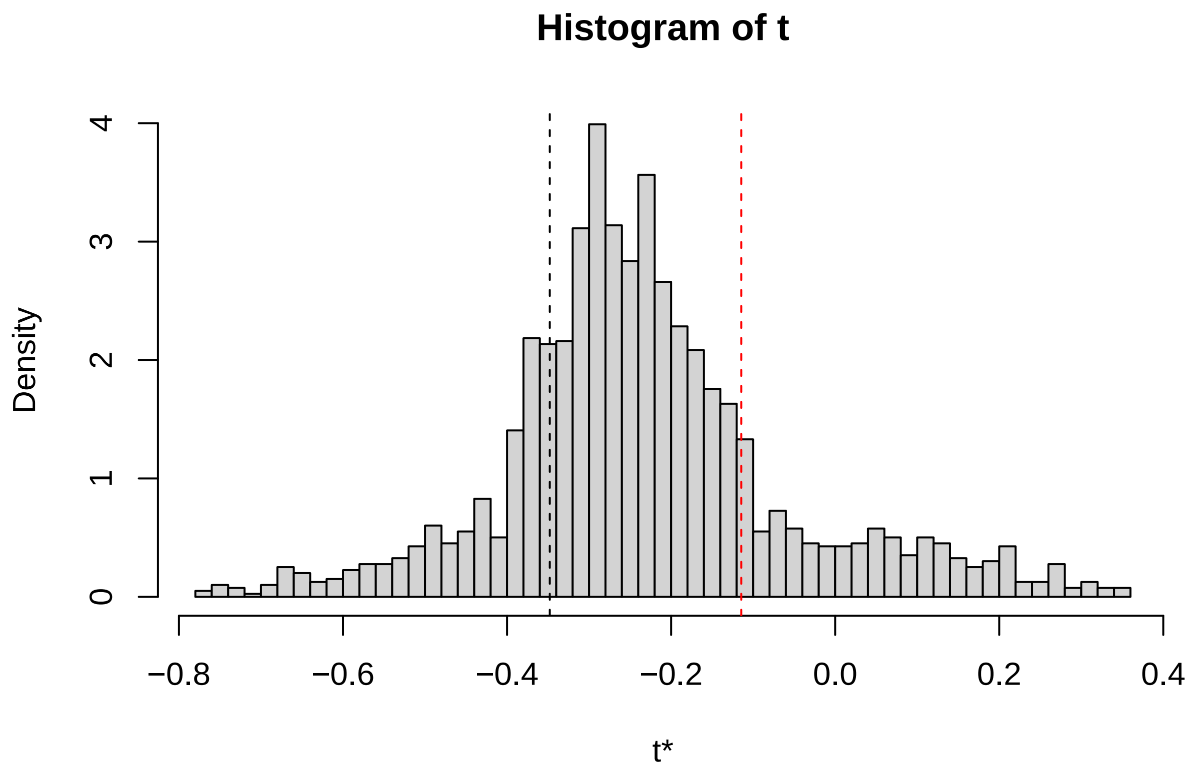
Lab 5

Group 4

2021-11-30

Question 1

```
lottery <- read.csv("lottery.csv", sep = ";")

t_test <- function(data, vn) {
  data <- data[vn,]
  fit <- loess(Draft_No ~ Day_of_year, data = lottery)
  Y_hat <- fit$fitted
  X_a <- data$Day_of_year[which.min(Y_hat)]
  X_b <- data$Day_of_year[which.max(Y_hat)]
  Y_hat_X_a <- Y_hat[X_a]
  Y_hat_X_b <- Y_hat[X_b]
  T_statistic <- (Y_hat_X_b - Y_hat_X_a) / (X_b - X_a)
  return(T_statistic)
}

# Non-parametric bootstrap
library(boot)

b_sample <- boot(lottery, t_test, R = 2000)

b_sample_ci <- boot.ci(b_sample, conf = 0.95, type = "basic")

hist(b_sample$t, breaks = 80, prob = TRUE, xlab = "t*", main = "Histogram of t")
abline(v = b_sample$t0, col = "black", lwd = 1, lty = 2)
abline(v = b_sample_ci$basic[4], col = "red", lwd = 1, lty = 2)
abline(v = b_sample_ci$basic[5], col = "red", lwd = 1, lty = 2)
```

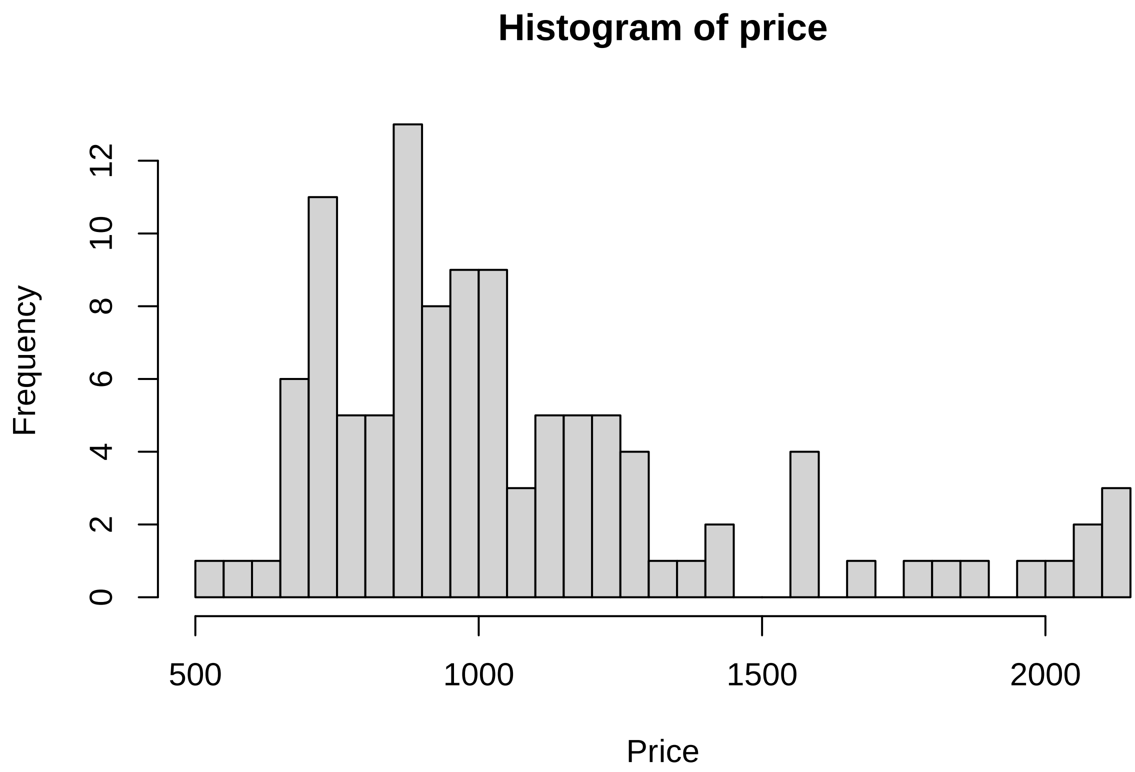


The expected value of T appear to be significantly different from 0. This means that the lottery is not random. The dashed, red line represents the upper limit of a 95% confidence interval for the expected value of T, which confirms that the expected value of T is in fact significantly different from 0 at a significance level of 0.05.

Question 2

2.1

```r
prices1 <- read.csv("prices1.csv", sep = ";")

hist(prices1$Price, breaks = 30, xlab = "Price", main = "Histogram of price")
```




```
mean(prices1$Price)
```

```
## [1] 1080.473
```

The shape of the distribution of prices looks similar to a log-normal distribution, or alternatively a gamma distribution.

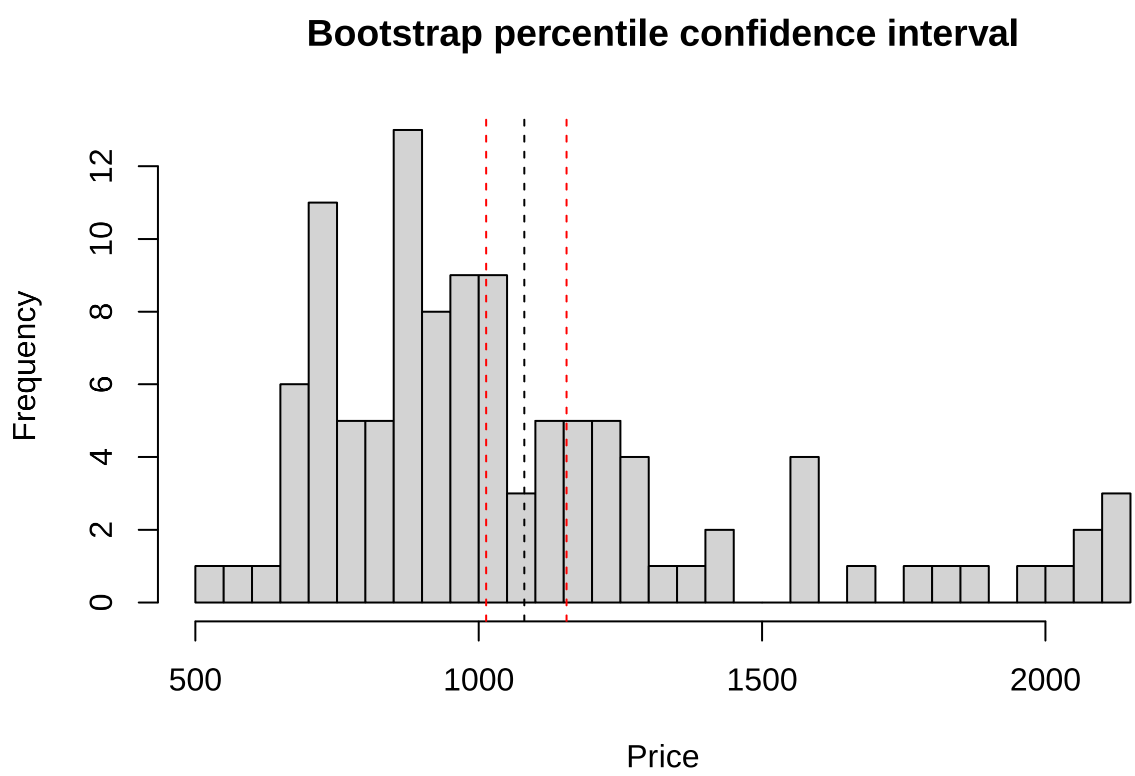
2.2

```
price_stat <- function(data, vn) {
  data <- data[vn,]
  return(mean(data$Price))
}

b_sample_2 <- boot(prices1, price_stat, R = 2000)

T1 <- 2 * mean(prices1$Price) - (1/length(b_sample_2$t) * sum(b_sample_2$t))

b_var <- (1/(length(b_sample_2$t))) * sum((b_sample_2$t - mean(b_sample_2$t))^2)

ci_perc <- boot.ci(b_sample_2, conf = 0.95, type = "perc")

ci_bca <- boot.ci(b_sample_2, conf = 0.95, type = "bca")
```

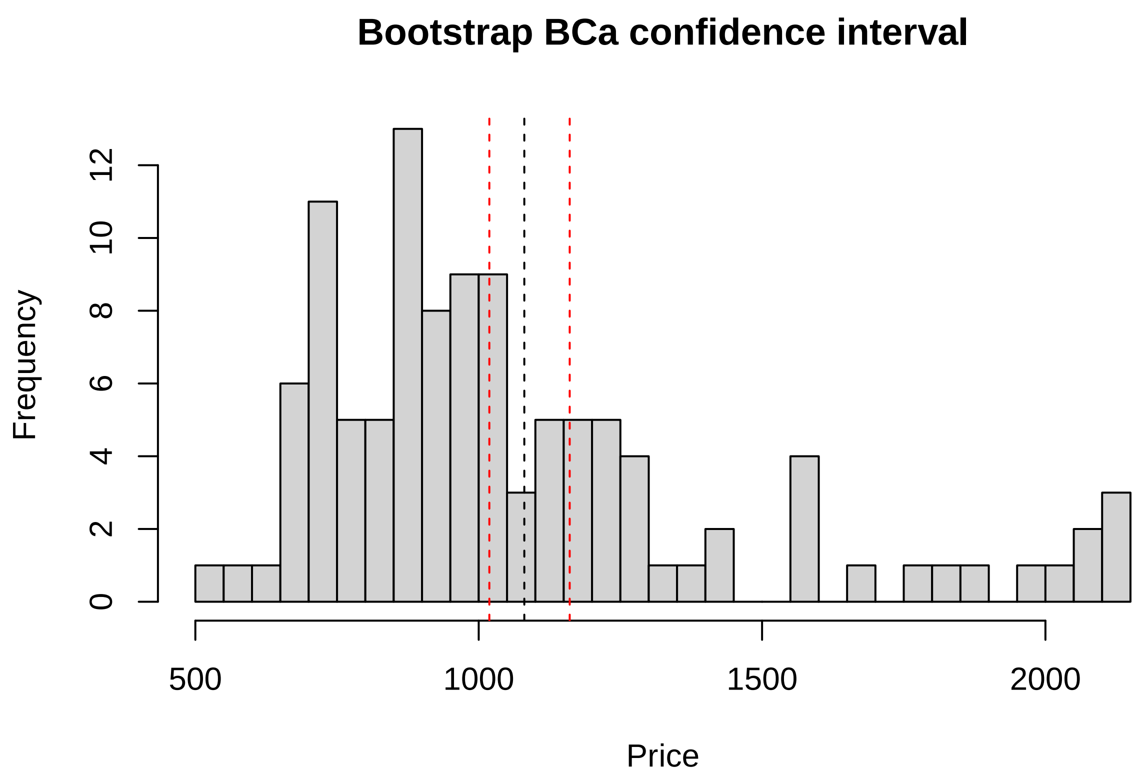
```
ci_norm <- boot.ci(b_sample_2, conf = 0.95, type = "norm")
```

2.3

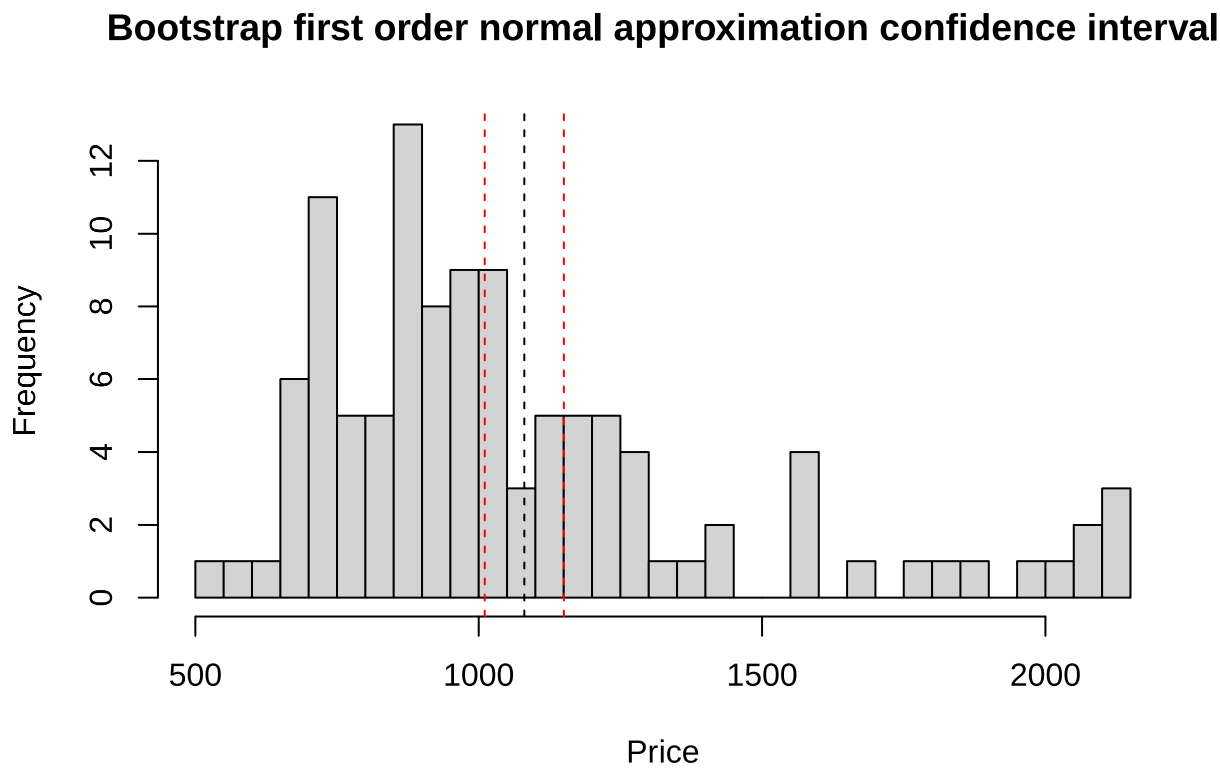
```
hist(prices1$Price, breaks = 30, xlab = "Price",
     main = "Bootstrap percentile confidence interval")
abline(v = ci_perc$percent[4], col = "red", lwd = 1, lty = 2)
abline(v = ci_perc$percent[5], col = "red", lwd = 1, lty = 2)
abline(v = T1, col = "black", lwd = 1, lty = 2)
```




```
hist(prices1$Price, breaks = 30, xlab = "Price",
     main = "Bootstrap BCa confidence interval")
abline(v = ci_bca$bca[4], col = "red", lwd = 1, lty = 2)
abline(v = ci_bca$bca[5], col = "red", lwd = 1, lty = 2)
abline(v = T1, col = "black", lwd = 1, lty = 2)
```




```r
hist(prices1$Price, breaks = 30, xlab = "Price",
     main = "Bootstrap first order normal approximation confidence interval")
abline(v = ci_norm$normal[2], col = "red", lwd = 1, lty = 2)
abline(v = ci_norm$normal[3], col = "red", lwd = 1, lty = 2)
abline(v = T1, col = "black", lwd = 1, lty = 2)
```



The black, dashed lines in the histograms shows the estimated mean price, and the dashed red lines shows the upper and lower limits of the confidence intervals. The estimated mean price is located within all of the three confidence intervals, which means that the estimate is significant at a 5% significance level.

2.4

```
# Jackknife estimate of the variance of the mean price
n <- length(prices1$Price)

T_D_i <- rep(0, n)

for(i in 1:n) {
  T_D_i[i] <- mean(prices1$Price[-i])
}

T_i <- n * mean(prices1$Price) - (n - 1) * T_D_i
J_T <- mean(T_i)

var_T_j <- (1/(n*(n-1))) * sum((T_i - J_T)^2)
var_T_j
```

```
## [1] 1320.911
```

```r
# Bootstrap estimate of the variance of the mean price
b_var
```

```
## [1] 1268.503
```

The variance of the mean price estimated by jackknife is slightly higher than the variance of the mean price estimated by bootstrapping. The jackknife variance estimator is known to often overestimate the variance (Gentle 2009).

References

Gentle, J. E. (2009). *Computational Statistics*. Springer.