# Untitled

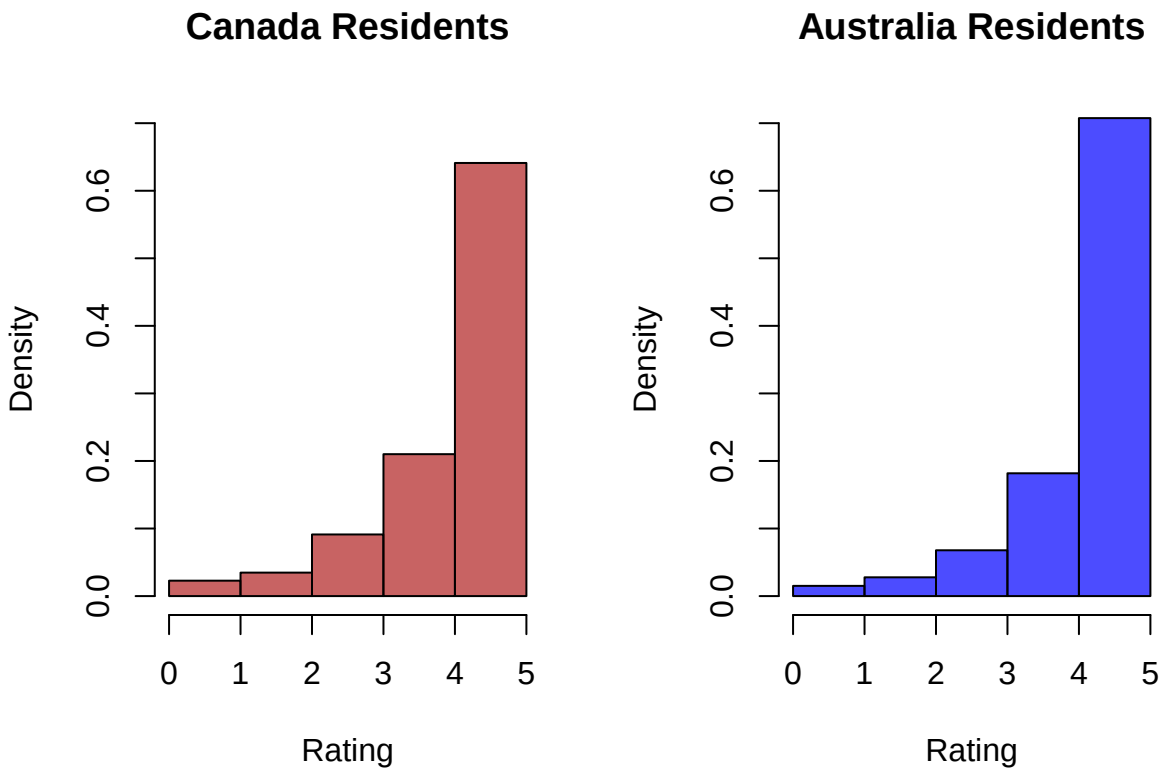
## QUESTION 1: Multiple Testing

(a) Construct two *density* histograms: one of the `Disneyland_California Rating` values made by residents of `Canada`, and the other of `Disneyland_California Rating` values made by residents of `Australia`. Plot them next to each other in a $1 \times 2$ grid, and include informative titles and axis labels. To enhance comparability ensure `breaks = 0:5` for both, and ensure `ylim` is the same for both.

```
disney <- read.csv("disneyland.csv")

pop <- list(pop1 = disney$Rating[disney$Reviewer_Location == "Canada" &
                                 disney$Branch == "Disneyland_California"],
            pop2 = disney$Rating[disney$Reviewer_Location == "Australia" &
                                 disney$Branch == "Disneyland_California"])

par(mfrow = c(1, 2))
hist(pop[[1]], col=adjustcolor("firebrick", 0.7), freq = FALSE,
     xlab = "Rating", main = "Canada Residents", breaks = 0:5, ylim = c(0, 0.7))
hist(pop[[2]], col=adjustcolor("blue", 0.7), freq = FALSE,
     xlab = "Rating", main = "Australia Residents", breaks = 0:5, ylim = c(0, 0.7))
```




(b) State the null hypothesis $H_0$ that is being tested when comparing these two sub-populations with a permutation test.

$H_0 : \mathcal{P}_1 and \mathcal{P}_2$ are both drawn from the same population of ratings

(c) Create a function `sdN(y)` that, given an input array `y`, calculates the population standard deviation

$$SD(\mathcal{P}) = \sqrt{\frac{1}{N}\sum_{u\in\mathcal{P}}(y_u - \overline{y})^2}.$$

Use your function to calculate the standard deviation of `y = c(4,9,3,2,7)`.

```
sdN <- function(y) {
  ybar = mean(y)
  sqrt(mean((y-ybar)^2))
}
sdN_test <- sdN(c(4,9,3,2,7))
print(sdN_test)
```

```
## [1] 2.607681
```

(d) [1 point] Create a function `skew(y)` that, given an input array `y`, calculates the population skewness

$$SKEW(\mathcal{P}) = \frac{\frac{1}{N}\sum_{u\in\mathcal{P}}(y_u - \overline{y})^3}{[SD(\mathcal{P})]^3}.$$

Use your function to calculate the skewness of `y = c(4,9,3,2,7)`.

```
skew <- function(y) {
  ybar = mean(y)
  mean((y-ybar)^3) / sdN(y)^3
}
skew_test <- skew(c(4,9,3,2,7))
print(skew_test)
```

```
## [1] 0.4060403
```

(e) [1 point] Create a function `kurt(y)` that, given an input array `y`, calculates the population kurtosis

$$KURT(\mathcal{P}) = \frac{\frac{1}{N}\sum_{u\in\mathcal{P}}(y_u - \overline{y})^4}{[SD(\mathcal{P})]^4}.$$

Use your function to calculate the kurtosis of `y = c(4,9,3,2,7)`.

```
kurt <- function(y) {
  ybar = mean(y)
  mean((y-ybar)^4) / sdN(y)^4
}
kurt_test <- kurt(c(4,9,3,2,7))
print(kurt_test)
```

```
## [1] 1.600346
```

(f) In this question you will test the hypothesis in (b) using the four discrepancy measures defined below, which respectively compare measures of center, spread, skewness, and kurtosis.

```
p_discrepancies <- list(D1, D2, D3, D4)
cache = TRUE
pval <- calculatePVmulti(pop, p_discrepancies, M_outer = 100, M_inner = 100)
print(pval) # p-val calculated is 0.02
```

```
## [1] 0.02
```

(g) Based on $p$-value$^\star$ calculated in (f), what do you conclude about the comparability of these two populations? In other words, summarize your findings and draw a conclusion about the null hypothesis from part (b). By referring to the histograms constructed in part (a), explain why this conclusion is, or is not, surprising.

Since the p-value is 0.02 which is very small, there is strong evidence against the hypothesis that the Canadian and Australian visitor ratings were randomly drawn from the same population of ratings based on the combined four discrepancies

The histograms are very similar as both are skewed to the right which suggests the population of ratings are also skewed to the right, so it was expected that the Canadian and Australian visitor ratings would likely be randomly drawn from the same population and it is suprising that this conclusion occurred

(h) Briefly explain why the approach taken in part (f) is to be preferred to considering four separate tests based on $D_1(\mathcal{P}_1, \mathcal{P}_2)$, $D_2(\mathcal{P}_1, \mathcal{P}_2)$, $D_3(\mathcal{P}_1, \mathcal{P}_2)$, and $D_4(\mathcal{P}_1, \mathcal{P}_2)$ individually.

The approach in part (f) considers multiple discrepancies into the test so it will be more robust compared to seperate tests.

## QUESTION 2: Bootstrap Confidence Intervals

(a) Subset the data found in `disneyland.csv` and retain only the reviews made by reviewers living in `Canada`. Include reviews for all three of the branches. This sub-population $\mathcal{P}$ should contain $N = 2235$ reviews. Using the indices contained in the `sampIndex.txt` file, take a sample $\mathcal{S}$ of size $n = 100$ from this sub-population. Print out `summary(Rating)` for this sample.

```
canada <- subset(disney, disney$Reviewer_Location == "Canada")$Rating
sampIndex <- read.table("sampIndex.txt")$V1
S <- canada[sampIndex]
summary(S)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    4.00    5.00    4.28    5.00    5.00
```

(b) By resampling $S$ with replacement, construct $B = 1000$ bootstrap samples $S^\star_1, S^\star_2, \ldots, S^\star_{1000}$.

```
B <- 1000
n <- 100
Sstar <- sapply(1:B, FUN = function(b) {
  sample(S, n, replace = TRUE)
})
```

(c) This question concerns the average `Rating`:

$$a(\mathcal{P}) = \overline{y} = \frac{1}{N} \sum_{u \in \mathcal{P}} y_u.$$

i. Calculate $a(\mathcal{S})$ and $a(\mathcal{P})$.

```
as <- mean(S) # calculate a(S)
as

ap <- mean(canada) # calculate a(P)
ap
```

```
## [1] 4.28
```

```
## [1] 4.298881
```

$$a(\mathcal{S}) = 4.28$$

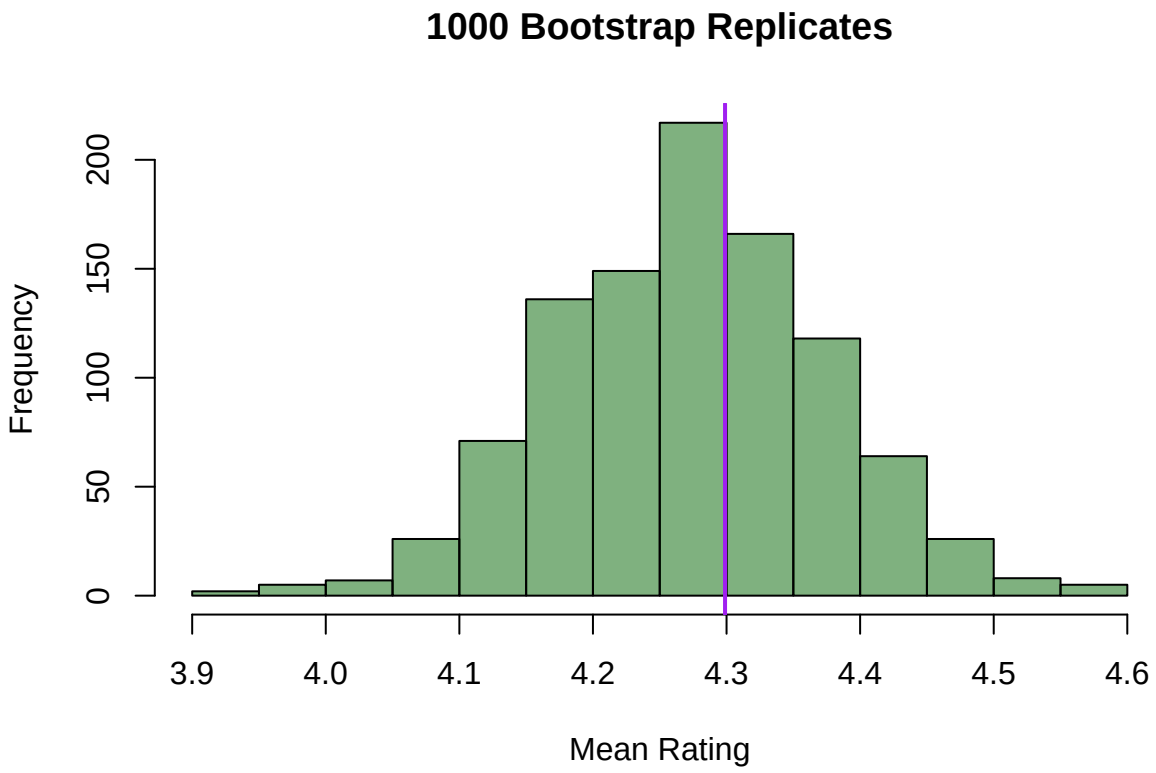
and

$$a(\mathcal{P}) = 4.298881$$

ii. Calculate $a(\mathcal{S}^\star_b)$ for each bootstrap sample $b = 1, 2, \ldots, B$ from part (b) and construct a histogram of these values. Be sure to include a vertical line representing $a(\mathcal{P})$, and also be sure to informatively label your plot.

```
as_star <- apply(X = Sstar, MARGIN = 2, FUN = mean)
hist(as_star, col = adjustcolor("darkgreen", 0.5), xlab = "Mean Rating",
```

```
     main = "1000 Bootstrap Replicates")
abline(v = ap, col = "purple", lwd = 2)
```




iii. Calculate a 95% confidence interval for $a(\mathcal{P})$ using the naive normal theory approach.

```
a_ci_naive <- as + 1.96 * c(-1, 1) * sd(as_star)
a_ci_naive
```

```
## [1] 4.080308 4.479692
```

iv. Calculate a 95% confidence interval for $a(\mathcal{P})$ using the percentile method.

```
a_ci_per <- c(quantile(as_star, 0.025), quantile(as_star, 0.975))
a_ci_per
```

```
##    2.5%   97.5%
## 4.08975 4.48000
```

v. Calculate a 95% confidence interval for $a(\mathcal{P})$ using the bootstrap-$t$ approach. **Note** that you may find it helpful to use the `bootstrap_t_interval` function defined on page 10. Please use $B = 1000$ and $D = 100$.

```
a_ci_boot <- bootstrap_t_interval(S = S, a = mean, confidence = 0.95,
                                  B = 1000, D = 100)
a_ci_boot
```

```
##    lower   middle    upper
## 4.061403 4.280000 4.479382
```

(d) This question concerns the standard deviation of `Rating`:

$$SD(\mathcal{P}) = \sqrt{\frac{1}{N}\sum_{u\in\mathcal{P}}(y_u - \overline{y})^2}.$$

i. Calculate $SD(\mathcal{S})$ and $SD(\mathcal{P})$. You may find your `sdN(y)` function from question 1(c) useful.

```
# sdN function from question 1
sds <- sdN(S)
sds

sdP <- sdN(canada)
sdP
```

```
## [1] 1.010742
```

```
## [1] 1.016276
```

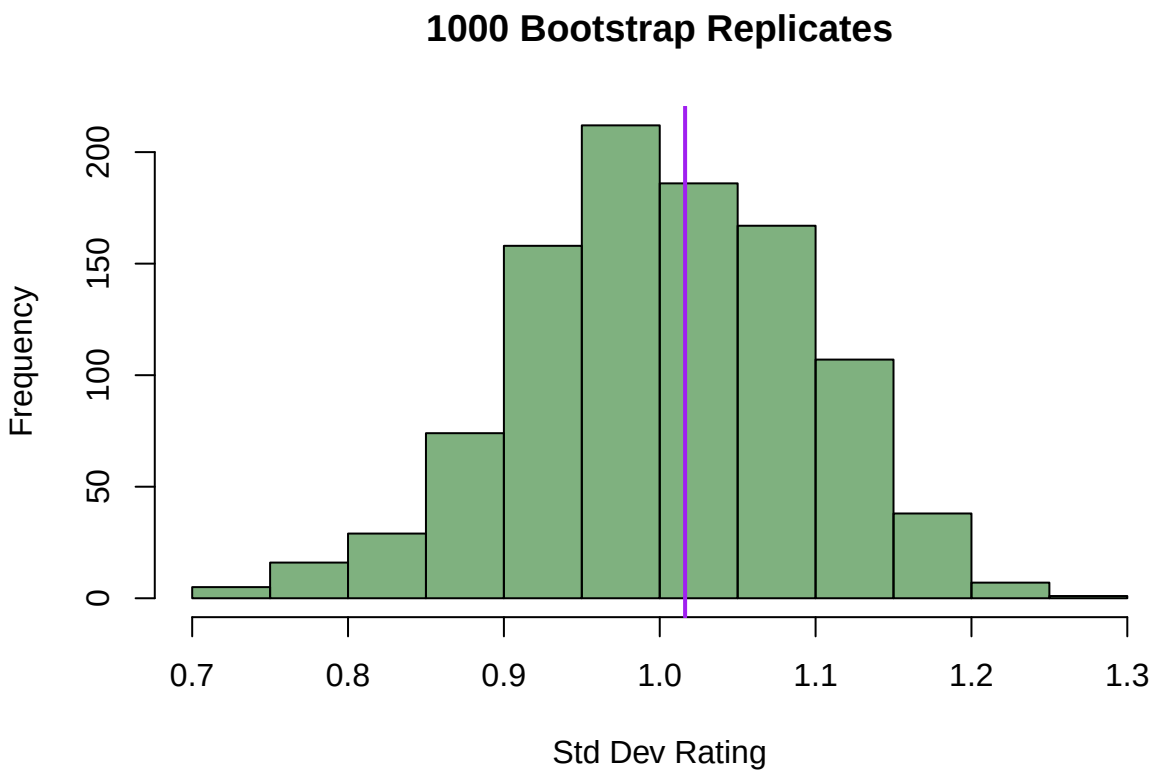
$$SD(\mathcal{S}) = 1.010742$$

and

$$SD(\mathcal{P}) = 1.016276$$

ii. Calculate $SD(\mathcal{S}_b^\star)$ for each bootstrap sample $b = 1, 2, \ldots, B$ from part (b) and construct a histogram of these values. Be sure to include a vertical line representing $SD(\mathcal{P})$, and also be sure to informatively label your plot.

```
sds_star <- apply(X = Sstar, MARGIN = 2, FUN = sdN)
hist(sds_star, col = adjustcolor("darkgreen", 0.5), xlab = "Std Dev Rating",
     main = "1000 Bootstrap Replicates")
abline(v = sdP, col = "purple", lwd = 2)
```




iii. Calculate a 95% confidence interval for $SD(\mathcal{P})$ using the naive normal theory approach.

```
sd_ci_naive <- sds + 1.96 * c(-1, 1) * sd(sds_star)
sd_ci_naive
```

```
## [1] 0.8312671 1.1902175
```

iv. Calculate a 95% confidence interval for $SD(\mathcal{P})$ using the percentile method.

```
sd_ci_per <- c(quantile(sds_star, 0.025), quantile(sds_star, 0.975))
sd_ci_per
```

```
##     2.5%    97.5%
## 0.806201 1.172348
```

v. Calculate a 95% confidence interval for $SD(\mathcal{P})$ using the bootstrap-$t$ approach. **Note** that you may find it helpful to use the `bootstrap_t_interval` function defined on page 10. Please use $B = 1000$ and $D = 100$.

```
sd_ci_boot <- bootstrap_t_interval(S = S, a = sdN, confidence = 0.95,
                                   B = 1000, D = 100)
sd_ci_boot
```

```
##     lower    middle     upper
## 0.8464243 1.0107423 1.2530572
```

(e) This question concerns the skewness of `Rating`:

$$SKEW(\mathcal{P}) = \frac{\frac{1}{N}\sum_{u\in\mathcal{P}}(y_u - \overline{y})^3}{[SD(\mathcal{P})]^3}.$$

i. Calculate $SKEW(\mathcal{S})$ and $SKEW(\mathcal{P})$. You may find your `skew(y)` function from question 1(d) useful.

```
# skew function from question 1
skews <- skew(S)
skews

skewP <- skew(canada)
skewP
```

```
## [1] -1.394666
```

```
## [1] -1.513027
```

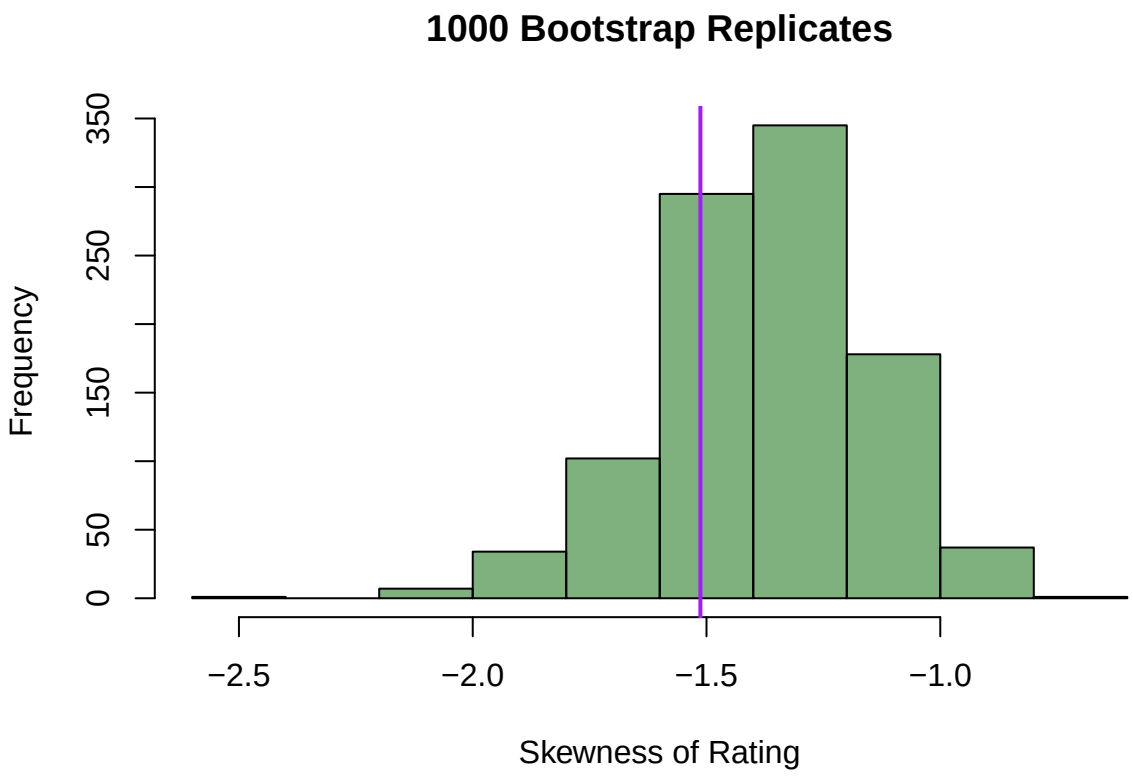
$$SKEW(\mathcal{S}) = -1.394666$$

and

$$SKEW(\mathcal{P}) = -1.513027$$

ii. Calculate $SKEW(\mathcal{S}_b^\star)$ for each bootstrap sample $b = 1, 2, \ldots, B$ from part (b) and construct a histogram of these values. Be sure to include a vertical line representing $SKEW(\mathcal{P})$, and also be sure to informatively label your plot.

```
skews_star <- apply(X = Sstar, MARGIN = 2, FUN = skew)
hist(skews_star, col = adjustcolor("darkgreen", 0.5), xlab = "Skewness of Rating",
     main = "1000 Bootstrap Replicates")
abline(v = skewP, col = "purple", lwd = 2)
```




iii. Calculate a 95% confidence interval for $SKEW(\mathcal{P})$ using the naive normal theory approach.

```
skew_ci_naive <- skews + 1.96 * c(-1, 1) * sd(skews_star)
skew_ci_naive
```

```
## [1] -1.8403815 -0.9489499
```

iv. Calculate a 95% confidence interval for $SKEW(\mathcal{P})$ using the percentile method.

```
skew_ci_per <- c(quantile(skews_star, 0.025), quantile(skews_star, 0.975))
skew_ci_per
```

```
##       2.5%      97.5%
## -1.8524689 -0.9823068
```

v. Calculate a 95% confidence interval for $SKEW(\mathcal{P})$ using the bootstrap-$t$ approach. **Note** that you may find it helpful to use the `bootstrap_t_interval` function defined on page 10. Please use $B = 1000$ and $D = 100$.

```
skew_ci_boot <- bootstrap_t_interval(S = S, a = skew, confidence = 0.95,
                                     B = 1000, D = 100)
skew_ci_boot
```

```
##     lower    middle     upper
## -1.956679 -1.394666 -1.008316
```

(f) This question concerns the kurtosis of `Rating`:

$$KURT(\mathcal{P}) = \frac{\frac{1}{N}\sum_{u\in\mathcal{P}}(y_u - \overline{y})^4}{[SD(\mathcal{P})]^4}.$$

i. Calculate $KURT(\mathcal{S})$ and $KURT(\mathcal{P})$. You may find your `kurt(y)` function from question 1(e) useful.

```
# kurt function from question 1
kurts <- kurt(S)
kurts

kurtP <- kurt(canada)
kurtP
```

```
## [1] 4.202684
```

```
## [1] 4.671102
```

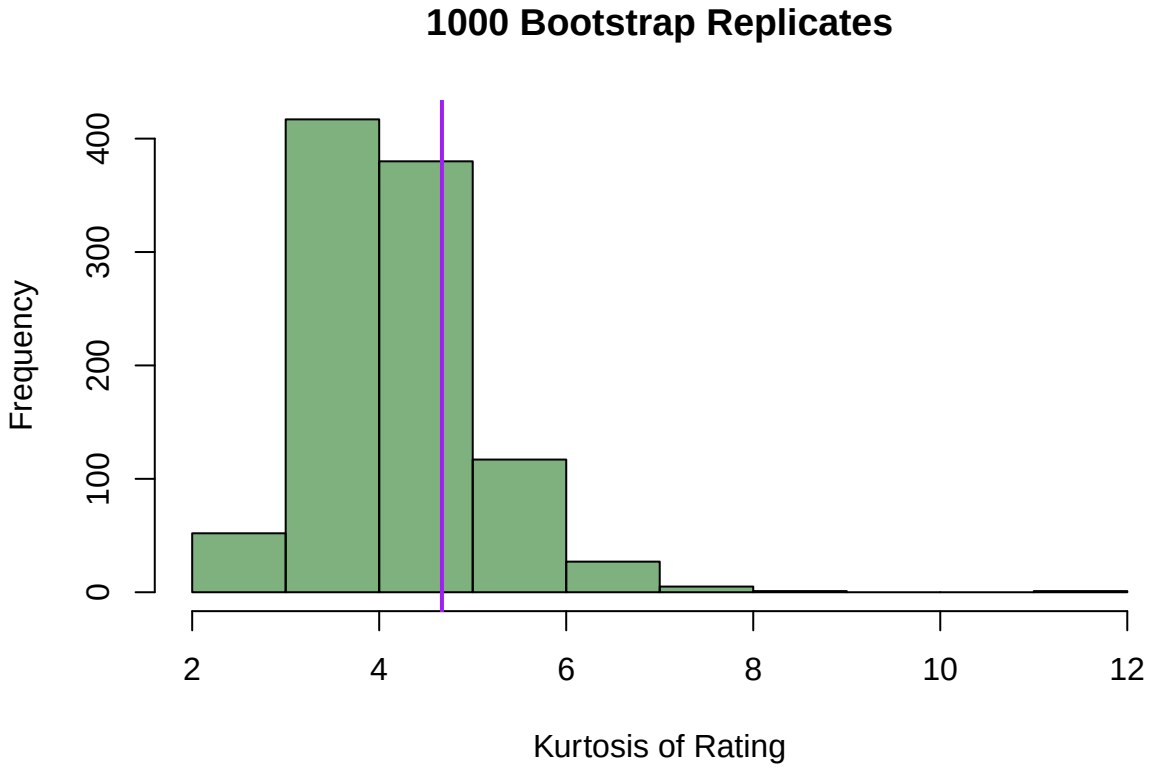
$$KURT(\mathcal{S}) = 4.202684$$

and

$$KURT(\mathcal{P}) = 4.671102$$

ii. Calculate $KURT(\mathcal{S}_b^\star)$ for each bootstrap sample $b = 1, 2, \ldots, B$ from part (b) and construct a histogram of these values. Be sure to include a vertical line representing $KURT(\mathcal{P})$, and also be sure to informatively label your plot.

```
kurts_star <- apply(X = Sstar, MARGIN = 2, FUN = kurt)
hist(kurts_star, col = adjustcolor("darkgreen", 0.5), xlab = "Kurtosis of Rating",
     main = "1000 Bootstrap Replicates")
abline(v = kurtP, col = "purple", lwd = 2)
```




iii. Calculate a 95% confidence interval for $KURT(\mathcal{P})$ using the naive normal theory approach.

```
kurt_ci_naive <- kurts + 1.96 * c(-1, 1) * sd(kurts_star)
kurt_ci_naive
```

```
## [1] 2.454608 5.950759
```

iv. Calculate a 95% confidence interval for $KURT(\mathcal{P})$ using the percentile method.

```
kurt_ci_per <- c(quantile(kurts_star, 0.025), quantile(kurts_star, 0.975))
kurt_ci_per
```

```
##     2.5%    97.5%
## 2.817345 6.328029
```

v. Calculate a 95% confidence interval for $KURT(\mathcal{P})$ using the bootstrap-$t$ approach. **Note** that you may find it helpful to use the `bootstrap_t_interval` function defined on page 10. Please use $B = 1000$ and $D = 100$.

```
kurt_ci_boot <- bootstrap_t_interval(S = S, a = kurt, confidence = 0.95,
                                     B = 1000, D = 100)
kurt_ci_boot
```

```
##    lower   middle    upper
## 2.925981 4.202684 6.812475
```

(g) This question concerns advantages and disadvantages associated with the various methods of confidence interval calculation you've explored.

i. List one advantage and one disadvantage of naive normal theory intervals.

One advantage is that it is straight forward to use when calculating confidence intervals.

One disadvantage is that it is accurate only if the bootstrap distribution is normal.

ii. List one advantage and one disadvantage of percentile method intervals.

One advantage is that it is straight forward to use when calculating confidence intervals.

One disadvantage is that it is mostly accurate if the distribution of the estimator is nearly symmetric.

iii. List one advantage and one disadvantage of bootstrap-$t$ intervals.

One advantage is that it automatically adjusts the shape of the sampling distribution estimator.

One disadvantage is that it is computationally intensive.

## QUESTION 3: Interview Questions

(a) In your own words, explain what it means to be "95% confident" in a 95% confidence interval. Aim your response at a level appropriate for the Senior Data Scientist who is interviewing you.

For random intervals, to determine an interval that contains the true population attribute, we can find a certain value $c$ such that the probability of the attribute's pivotal quantity being bounded between $-c$ and $c$ is equivalent to the coverage probability $1 - p$. Since confidence intervals are an observed version of a random interval, $1 - p$ is referred to as the confidence level for confidence intervals instead.

So in a 95% confidence interval, $1 - p = 0.95$ which represents a 95% confidence level and means we are 95% confident that the true population attribute is contained in the interval.

(b) In your own words, explain what resampling methods like the bootstrap are, and why they are useful. Aim your response at a level appropriate for the Senior Data Scientist who is interviewing you.

The main idea of resampling methods is that it uses a sample $S$ as if it were the population and repeatedly draw samples from it. More specifically, we draw $B$ samples of a certain size $n$ independently with replacement from the imitation population $S*$.

Resampling methods are useful because it uses a single sample which is used to construct an estimate of the sampling distribution of an attribute without assumptions about the form on the attribute, so it provides a bootstrap distribution which is a proxy for the sampling distribution for any sample attribute. That way we can construct confidence intervals for population attributes without worrying about the sampling distribution.