Bootstrap Confidence Intervals

Disneyland Resort opened in Anaheim California in 1955 and was the first of many Disney theme parks to have opened around the world. Three of these parks fall under the "Disneyland" branding: Disneyland Resort, Disneyland Paris, and Hong Kong Disneyland Resort. The file disneyland.csv contains information concerning Tripadvisor reviews of these three Disneyland branches. In particular, for each of 42,656 Tripadvisor reviews the various variates have been recorded, including rating, a numeric value indicating the reviewers satisfaction with their visit.

We will first subset the data and only include reviews made by reviewers living in Canada.

```
sdN = function(y) {
  sqrt(mean((y - mean(y))^2 ))
}
```

```
set.seed(341)

disney <- read.csv("disneyland.csv", header = TRUE)

canada <- subset(disney, disney$Reviewer_Location == "Canada")

n <- 100

N <- dim(canada)[1]

SampIndex <- read.table("sampIndex.txt")$V1
s <- canada[SampIndex, "Rating" ]

summary(s)
```

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

By resampling with replacement, we then construct 1000 bootstrap samples.

```
B <- 1000

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

Average

We see the population and sample average.

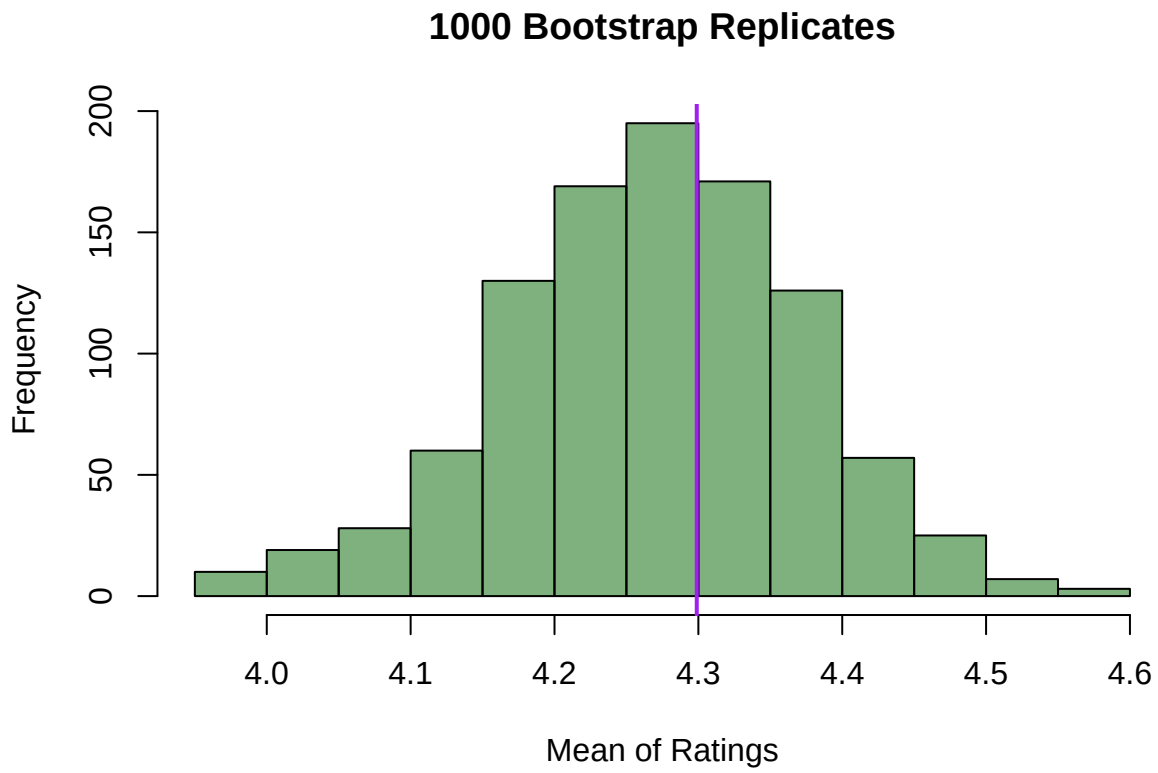
```
ap <- mean(canada$Rating)
as <- mean(s)
```

$a(\mathcal{P}) = 4.2988814$ and $a(\mathcal{S}) = 4.28$

We calculate the bootstrap average for each bootstrap sample, and construct a histogram of these values, with a vertical line representing the population average.

```
a_star <- apply(X = Sstar, MARGIN = 2, FUN = mean)

hist(a_star, col = adjustcolor("darkgreen", 0.5), xlab = "Mean of Ratings",
     main = "1000 Bootstrap Replicates")
abline(v = mean(canada$Rating), col = "purple", lwd = 2)
```



95% C.I of average with naive normal theory approach

```
set.seed(341)

mean(s) + qnorm(0.95) * c(-1, 1) * sdN(a_star)
```

```
## [1] 4.109107 4.450893
```

The 95% C.I using the naive normal theory approach is (4.109107, 4.450893).

95% C.I of average with percentile method.

```
set.seed(341)

c(quantile(a_star, 0.05), quantile(a_star, 0.95))
```

```
##   5%  95%
## 4.10 4.44
```

The 95% C.I using the percentile approach is (4.10, 4.44).

95% C.I of average with bootstrap-t approach.

```
bootstrap_t_interval <- function(S, a, confidence, B, D) {
## Inputs: S = an n element array containing the variate values in the
## sample a = a scalar-valued function that calculates the attribute a()
## of interest confidence = a value in (0,1) indicating the confidence
## level B = a numeric value representing the outer bootstrap count of
## replicates (used to calculate the lower and upper limits) D = a
## numeric value representing the inner bootstrap count of replicates
## (used to estimate the standard deviation of the sample attribute for
## each (outer) bootstrap sample)
Pstar <- S
aPstar <- a(Pstar)
sampleSize <- length(S)
## get (outer) bootstrap values
bVals <- sapply(1:B, FUN = function(b) {
Sstar <- sample(Pstar, sampleSize, replace = TRUE)
aSstar <- a(Sstar)
## get (inner) bootstrap values to estimate the SD
Pstarstar <- Sstar
SD_aSstar <- sd(sapply(1:D, FUN = function(d) {
Sstarstar <- sample(Pstarstar, sampleSize, replace = TRUE)
## return the attribute value
a(Sstarstar)
}))
z <- (aSstar - aPstar)/SD_aSstar
```

```
## Return the two values
c(aSstar = aSstar, z = z)
})
SDhat <- sd(bVals["aSstar", ])
zVals <- bVals["z", ]
## Now use these zVals to get the lower and upper c values.
cValues <- quantile(zVals, probs = c((1 - confidence)/2, (confidence +
1)/2), na.rm = TRUE)
cLower <- min(cValues)
cUpper <- max(cValues)
interval <- c(lower = aPstar - cUpper * SDhat, middle = aPstar, upper = aPstar -
cLower * SDhat)
return(interval)
}
```

```
set.seed(341)

bootstrap_t_interval(S=s, a=mean, confidence = 0.95, B = 1000, D = 100)
```

```
##    lower   middle    upper
## 4.037091 4.280000 4.470638
```

The 95% C.I using the bootstrap-t approach is (4.037091, 4.470638).

Standard Deviation

We see the population and sample standard deviation.

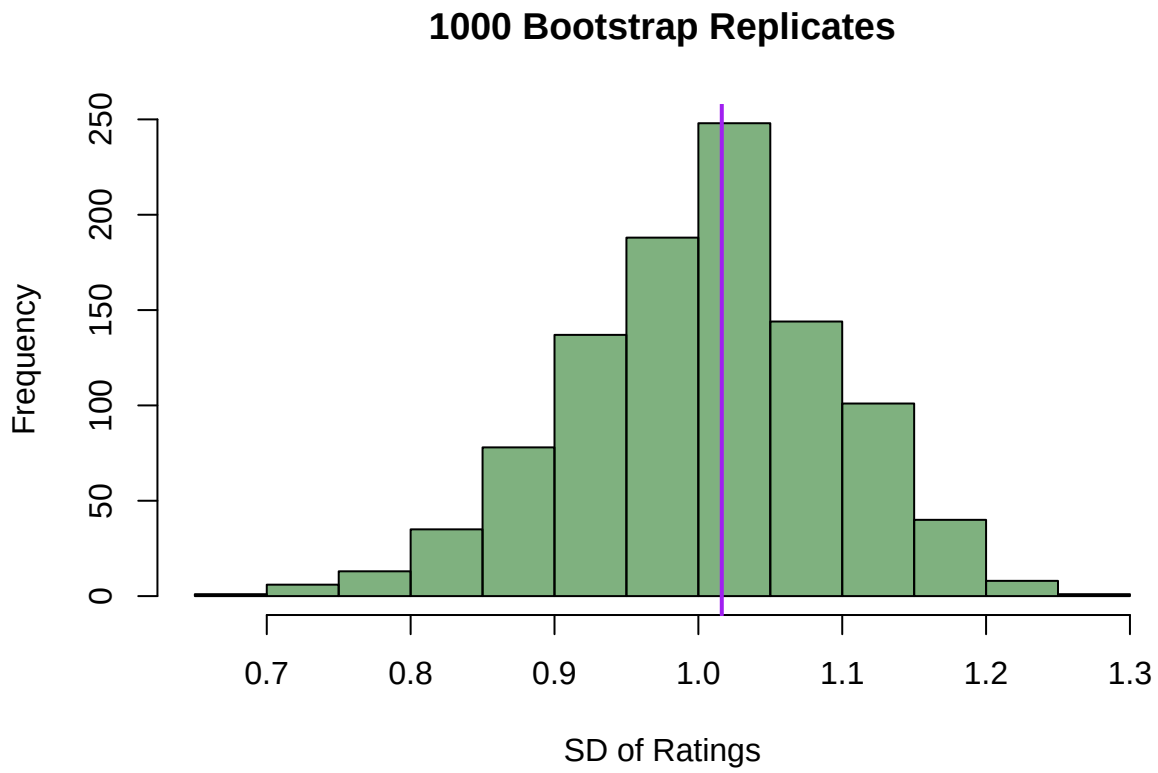
```
sp <- sdN(canada$Rating)
ss <- sdN(s)
```

$a(\mathcal{P}) = 1.0162763$ and $a(\mathcal{S}) = 1.0107423$

We calculate the bootstrap SD for each bootstrap sample, and construct a histogram of these values, with a vertical line representing the population SD.

```
s_star <- apply(X = Sstar, MARGIN = 2, FUN = sdN)

hist(s_star, col = adjustcolor("darkgreen", 0.5), xlab = "SD of Ratings",
     main = "1000 Bootstrap Replicates")
abline(v = sdN(canada$Rating), col = "purple", lwd = 2)
```



95% C.I of standard deviation with naive normal theory approach

```
set.seed(341)

sdN(s) + qnorm(0.95) * c(-1, 1) * sdN(s_star)
```

```
## [1] 0.8590696 1.1624150
```

The 95% C.I using the naive normal theory approach is (0.8590696, 1.1624150).

95% C.I of standard deviation with percentile method

```
set.seed(341)

c(quantile(s_star, 0.05), quantile(s_star, 0.95))
```

```
##        5%       95%
## 0.8443487 1.1465948
```

The 95% C.I using the percentile approach is (0.8443487, 1.1465948).

95% C.I of standard deviation with bootstrap-t approach

```
set.seed(341)

bootstrap_t_interval(S=s, a=sdN, confidence = 0.95, B = 1000, D = 100)
```

```
##     lower    middle     upper
## 0.8298959 1.0107423 1.2826772
```

The 95% C.I using the bootstrap-t approach is (0.8298959, 1.2826772).