CALIFORNIA STATE UNIVERSITY
EAST BAY

STAT 641- BOOTSTRAPPING METHODS

Spring Semester 2020.

PROJECT REPORT FOR

TENNIS & BOOTSTRAP CONFIDENCE INTERVALS.

Date: 03/09/20.

Submitted By:

Sajith Gowthaman (ek5282)

Anish Raviraj (ej4362)

Edgar Vidriales (xd9352)

Nima Razavi (ue7465)

SUBMITTED TO: Prof. Ayona Chatterjee

INTRODUCTION:

We have performed bootstrapping for the ATP world tour tournament data consisting of years 2018 and 2016. We will do this by finding the confidence intervals for the difference between the mean amount of aces that are hit between the winners and the losers of matches. Along with finding the differences in ace count between all winners and losers in 2018 for three set matches, we have also looked at the **differences in ace count based on surface played and looked at the differences in height between winners and losers of matches.** The 2018 ATP World Tour was the global elite men's professional tennis circuit organized by the Association of Tennis Professionals (ATP) for the 2018 tennis season. The 2018 ATP World Tour calendar comprises the Grand Slam tournaments (supervised by the International Tennis Federation (ITF)), the ATP World Tour Masters 1000, the ATP World Tour 500 series, the ATP World Tour 250 series, the Davis Cup (organized by the ITF), and the ATP Finals. We have collected the dataset for bootstrapping the ATP world tennis tournament for the year 2018 and 2016 from the data bank named "Kaggle". This project will help the statistician, coaches of Tennis sport to gain insights and will help predict the most likely outcome of the game through this analysis.

DIFFERENT TYPES OF COURTS:

We have considered a total of three types of court:

(i) Hard court
(ii) Clay Court
(iii) Grass Court.

Hard Court:

Hard courts can range from faster to slower speeds depending on the quantity and size of sand mixed into the paint coating. Hard courts tend to equalize the playing field in terms of athletic style. An acrylic hard court is used in the US Open and a synthetic for the Australian Open. While acrylic courts are more rigid and create a faster game, they can also be rough on the human body. To combat this, artificial tennis court surfaces were created to allow for similar usability and low maintenance, but also allow for greater shock absorption for players. These courts have been installed everywhere from the White House and the Sony Ericsson Open to high schools and tennis clubs around the world.

Of the Grand Slam tournaments, the US Open and Australian Open currently use hard courts and it is the predominant surface type used on the professional tour.

Clay Court:

Clay courts are made of compressed shale, stone, or brick. While they are relatively easy and cheap to install, long-term maintenance can be quite expensive, particularly when compared to the costs of artificial tennis court surfaces. The water balance within the clay must be carefully monitored and adjusted, and the court must be rolled periodically to preserve its flatness.
In terms of play, clay courts are traditionally used in the French Open. These courts characteristically have a slower game and give balls a higher bounce. They tend to favor baseline players, as well. These courts are most popular in Europe and Latin America.

Grass Court:

Tennis courts with grass surfaces are not as common today as they have been in the past due to the high maintenance costs of constant watering and mowing. They are also more impacted by weather conditions than clay or hard-court surfaces are. Dirt is hard-packed, and the grass must be trimmed very short. Grass is the fastest type of court because of its low bounce capacity. Players must get to the ball much more quickly than with clay or hard court. This means that players with stronger serve-and-volley skills will generally perform better. The grass court is the signature of Wimbledon.

The number of aces for each winner in 2018 is found by using the function in Rstudio.

```
Winner's ace count subsetting
```{r}
library(tidyverse)
WinnerACE18<-atp_matches_2018 %>% select(w_ace, best_of) %>% filter(best_of=="3") #dropping
5 set macthes
summary(WinnerACE18$w_ace)
WinnerACE18<-na.omit(WinnerACE18) #removes missing values

quantile(WinnerACE18$w_ace,c(.1, .9))

WinnerACE18<-filter(WinnerACE18, w_ace<16, w_ace>2) #trimming data
glimpse(WinnerACE18)

summary(WinnerACE18)

mean(WinnerACE18$w_ace)
sd(WinnerACE18$w_ace)
median(WinnerACE18$w_ace)
hist(WinnerACE18$w_ace)
abline(v=mean(WinnerACE18$w_ace), main= "Histogram of Winner's Ace Count For 2018")
```
```

In this, we are using the package “tidyverse” which allows us to subset the dataset to where we only have matches of **three** sets. When subsetting our data, all missing values were dropped. We also decided to trim the values because of outliers in the data set.

Firstly, we are look at our newly subsetted data. The histogram below shows the distribution of aces for the winner of each match. The histogram is skewed right with a large peak before four aces.

METHODOLOGY

The We have planned to go about is to find the ace count for winners and losers and do a bootstrap for it using a sample of 10,000 trials. Then we took a bootstrap for the difference between the winners and losers. Last part of the project involved the usage of library “tidyverse” where we took the ace count of winners and losers grouped by the surfaces. The results were plotted in the form of histogram and a 95% confidence interval is plotted to find the difference.

DISCUSSION:

WINNER VS LOSER ACE COUNT:

Next, we wanted to see if there is a difference between the average amount of aces that were hit by the winner of the match and average amount of aces hit by the loser of the match. We bootstrapped a sample of 10,000 and for-looped the mean difference of the sample. Then the histogram was plotted for visualizing the normality and where the 95% confidence intervals lie. An “ab” line is added in addition showing the point where the difference between the means occur.

```
Bootstrap for difference in aces between winner and loser
```{r}
B=10^4
diff_mean=numeric(B) #Same as using NULL
for(i in 1:B)
{
  w.sample=sample(WinnerACE18$w_ace, 1710, T)
  l.sample=sample(LoserACE18$l_ace, 1930, T)
  diff_mean[i]=mean(w.sample)- mean(l.sample)
}
hist(diff_mean, main = "Histogram of Bootstrap for Differences in Mean Between \nWinner's
and Loser's Ace Count in 2018")
abline(v=mean(WinnerACE18$w_ace)-mean(LoserACE18$l_ace), col="blue", lty=2)

quantile(diff_mean, c(0.025, 0.975)) #95% BS Percentile CI
```
```

Based on the bootstrap percentile confidence interval, we are 95% confident that, on average, winners have an ace count that is between **2.329** and **2.7** higher than the losers of the match. Since zero is not in our percentile confidence interval, we can support the claim that **there is a significant difference between the average amount of aces that the winner and loser of a match hit.**

HARD COURT VS GRASS COURT:

We are making a comparison for the number of aces served by the winner of the match on two different surfaces. First, we will look at the average ace count for the winner of a match on hard court and grass court. To begin, the following chunk was run to subset for hard courts. A visualizing using a histogram to get an understanding of the distribution of mean number of aces for the winner on the hard court was made. Just like the histogram for all the surfaces, there is a right skew in our data. Similar to the above bootstrap, we have used the histogram to plot its mean, standard deviation, median and histogram.

```
Subsetting for hard court
```{r}
library(tidyverse)
surfaceACE18<-atp_matches_2018 %>% select(w_ace, best_of, surface)

hard18<-filter(surfaceACE18, best_of=="3", surface=="Hard")
hard18<-na.omit(hard18) #Gets rid of missing values

hist(hard18$w_ace) # still skewed, but not as skewed
summary(hard18$w_ace)

quantile(hard18$w_ace,c(.1, .9))

hard18<-filter(hard18, w_ace<15, w_ace>1)

summary(hard18)

mean(hard18$w_ace)
sd(hard18$w_ace)
median(hard18$w_ace)
hist(hard18$w_ace,main =  "Histogram of Winner's Ace Count for Hard Court") #now trimmed
abline(v=mean(hard18$w_ace))
```
```

Similarly, we are running the chunk for grass court and visualizing the distribution of aces served by the winner of the match.

```
Subsetting for grass
```{r}
grass18<-filter(surfaceACE18, best_of=="3", surface=="Grass")
grass18<-na.omit(hard18)

hist(grass18$w_ace)
summary(grass18$w_ace)

quantile(grass18$w_ace,c(.1, .9))

hard18<-filter(grass18, w_ace<17, w_ace>2)

summary(grass18)

mean(grass18$w_ace)
sd(grass18$w_ace)
median(grass18$w_ace)
hist(grass18$w_ace, main =  "Histogram of Winner's Ace Count for Grass Court") #now trimmed
abline(v=mean(grass18$w_ace))
```
```

Once we have subsetted the data for both hard and grass court, we ran a bootstrap for the difference in the average number of aces between the two courts. The bootstrap performed subtracted the mean amount of aces on grass court from the mean of aces on hard court during each resample. **Prior to running the bootstrap, we expected there to be higher amount of aces on grass court than on court since tennis tends to be played the fastest on grass.**

```
Hard court vs Grass court
```{r}
B=10^4
hg_diff_mean=numeric(B) #Same as using NULL
set.seed(5)
for(i in 1:B)
{
  hard.sample=sample(hard18$w_ace, 1359, T)
  grass.sample=sample(grass18$w_ace, 165, T)
  hg_diff_mean[i]=mean(hard.sample)- mean(grass.sample)
}
hist(hg_diff_mean, main="Histogram of Bootstrap for Difference in Means of Ace Count \nfor
the Winner on Hard Court vs. Grass Court")
abline(v=mean(hard18$w_ace)-mean(grass18$w_ace), col="blue", lty=2)

quantile(hg_diff_mean, c(0.025, 0.975)) #95% BS Percentile CI
```

After running 10,000 trials, we are 95% confident that, on average, **winner's on grass court have an ace count that is between 0.183 and 1.245 higher than winner's on hard courts do.** This result goes with along with our hypothesis that grass court winners have a higher ace count than hard court winners do.

HARD VS CLAY COURT:

We are running the chunk for clay court, to find the mean number of ace won by winners and losers in the clay court. Visualization is done by histogram to see the distribution of ace with clay court.
```

```
Subsetting for clay court
```{r}
clay18<-filter(surfaceACE18, best_of=="3", surface=="Clay")
clay18<-na.omit(clay18)

hist(clay18$w_ace)
summary(clay18$w_ace)

quantile(clay18$w_ace,c(.1, .9))

clay18<-filter(clay18, w_ace<15, w_ace>1)

summary(clay18)

mean(clay18$w_ace)
sd(clay18$w_ace)
median(clay18$w_ace)
hist(clay18$w_ace, main =  "Histogram of Winner's Ace Count for Clay Court") #now trimmed
abline(v=mean(clay18$w_ace))
```
```

Next, we bootstrapped a sample of 10,000 and took the mean of it by for-looping the difference in mean. Then the histogram was plotted for finding the confidence interval at 95%.

```
Hard court vs Clay court
```{r}

set.seed(2)
B=10^4
hc_diff_mean=numeric(B) #Same as using NULL
for(i in 1:B)
{
  hard.sample=sample(hard18$w_ace, 1359, T)
  clay.sample=sample(clay18$w_ace, 536, T)
  hc_diff_mean[i]=mean(hard.sample)- mean(clay.sample)
}
hist(hc_diff_mean, main="Histogram of Bootstrap for Difference in Means of Ace Count \nfor
the Winner on Hard Court vs. Clay Court")
abline(v=mean(hard18$w_ace)-mean(clay18$w_ace), col="blue", lty=2)

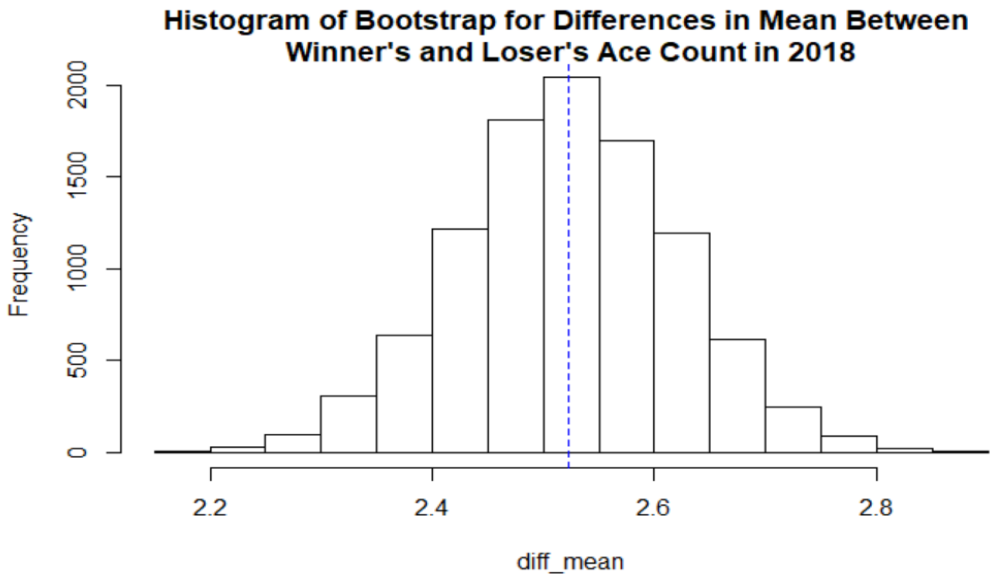
quantile(hc_diff_mean, c(0.025, 0.975)) #95% BS Percentile CI
```

We do have significance, as there is a difference between the mean amount of aces for the winner on hard court match and the winner on a clay court.

We are 95% confidence that, on average**, the winner of a hard court will have between 0.997 and 1.59 more aces than the winner of a clay court match. The rest were covered by the other team mates.**
```

RESULTS:

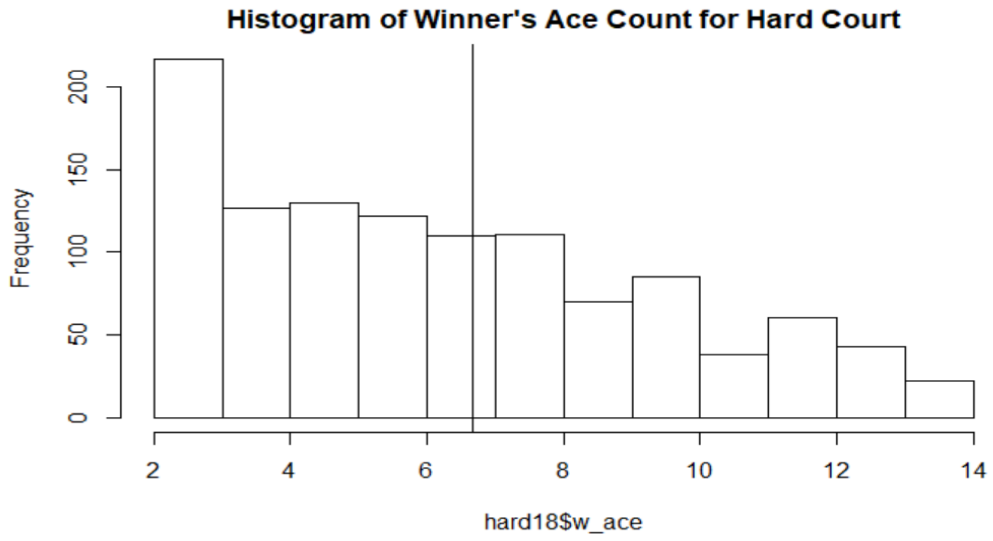
1) Mean difference between Winner's and Loser's Count:



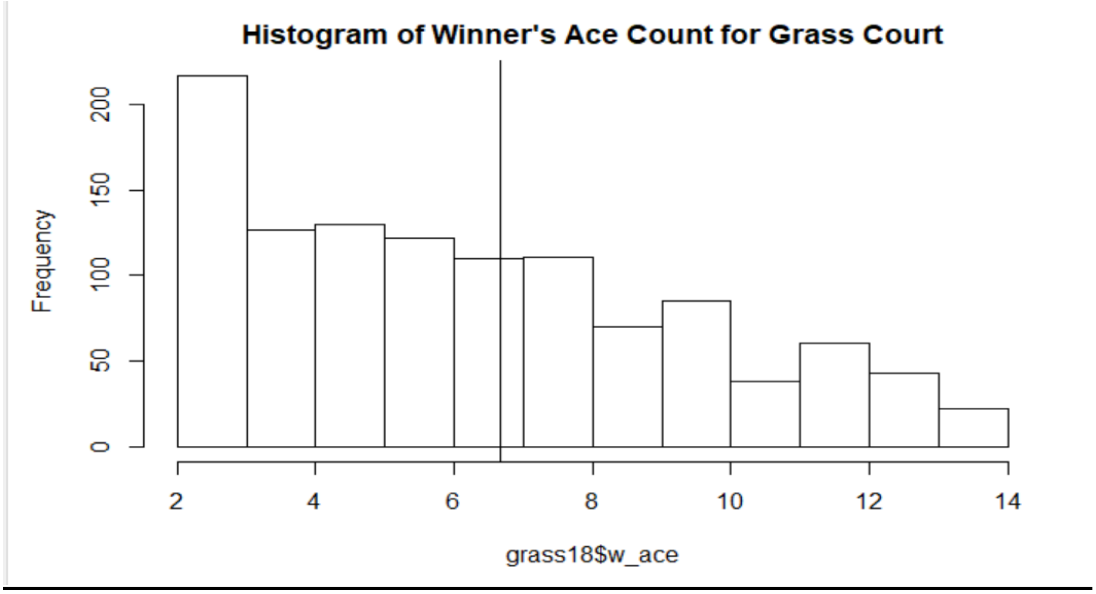
Winners have an ace count that is between **2.329** and **2.7** higher than the losers of the match. There is a Significant Difference.

2) Mean difference between Hard Court and Grass Court

Visualizing the distribution of aces served by the winner of the match on the Hard Court

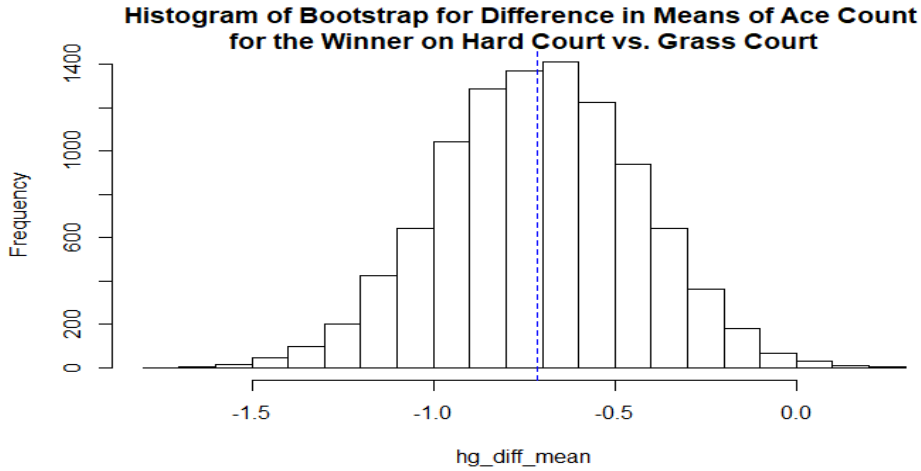


Visualizing the distribution of aces served by the winner of the match on the Grass Court

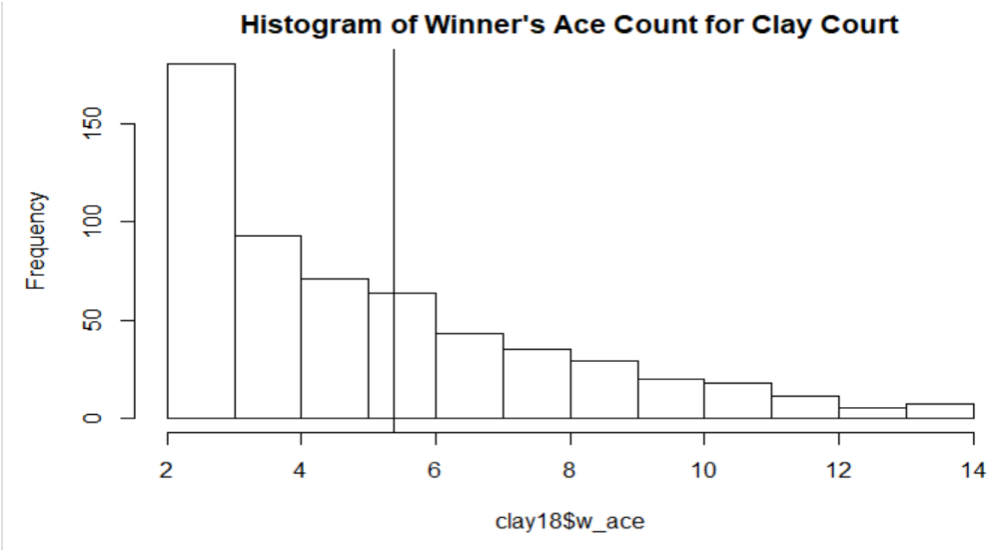


Difference between the winners for the Hard Court vs Grass Court

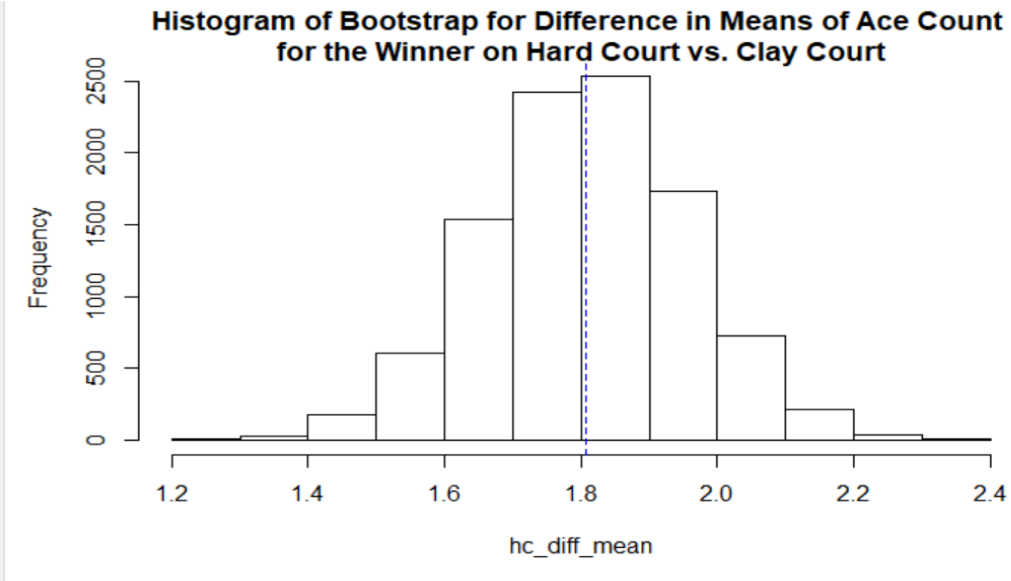
The **winners on grass court have an ace count that is between 0.183 and 1.245 higher than winner's on hard courts do.**



Winner's Ace count for Clay Court.



the winner of a hard court will have between 0.997 and 1.59 more aces than the winner of a clay court match.



CONCLUSION:

From the first where we took the difference between the means of winners and losers ace account we can conclude that the winners have a higher ace count compared to that of the losers and that there is a difference in the averages of the amount of aces between winners and losers. From the second plot between hard court and grass court, it is evident that the winners on the grass court have ace counts higher compared to that of the winners on the hard court. The last bootstrapped plot tells us that the winners on the hard court will be more than that of the clay court.

APPENDICES:

Describes above along with its explanation and plots.

REFERENCES:

Dataset of ATP tennis tournament 2018 to conduct bootstrapping:

https://www.kaggle.com/pablodroca/atp-tennis-matches-20002019