# 466 Final project_1

JongHee Lee

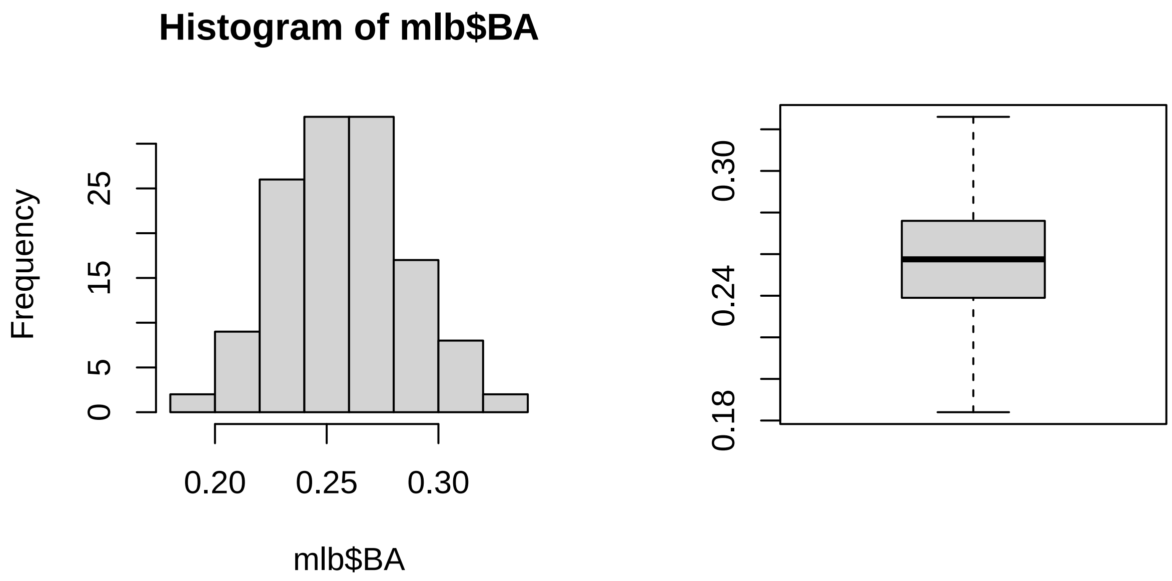
2023-04-14

In this project, we aim to predict the population mean of Batting Averages (BA) for Major League Baseball (MLB) players in the 2022 season, using player-specific statistics as our foundation. In the sport of baseball, BA is an essential metric that represents a batter's ability to hit the ball successfully, calculated as the number of hits divided by the number of at-bats.

```
head(mlb)
```

```
## # A tibble: 6 x 4
##   Name                  Age Tm       BA
##   <chr>               <dbl> <chr> <dbl>
## 1 Jeff McNeil*           30 NYM   0.326
## 2 Freddie Freeman*       32 LAD   0.325
## 3 Paul Goldschmidt       34 STL   0.317
## 4 Luis Arraez*           25 MIN   0.316
## 5 Aaron Judge            30 NYY   0.311
## 6 Xander Bogaerts        29 BOS   0.307
```

```
par(mfrow=c(1,2))
hist(mlb$BA)
boxplot(mlb$BA)
```



# EDA

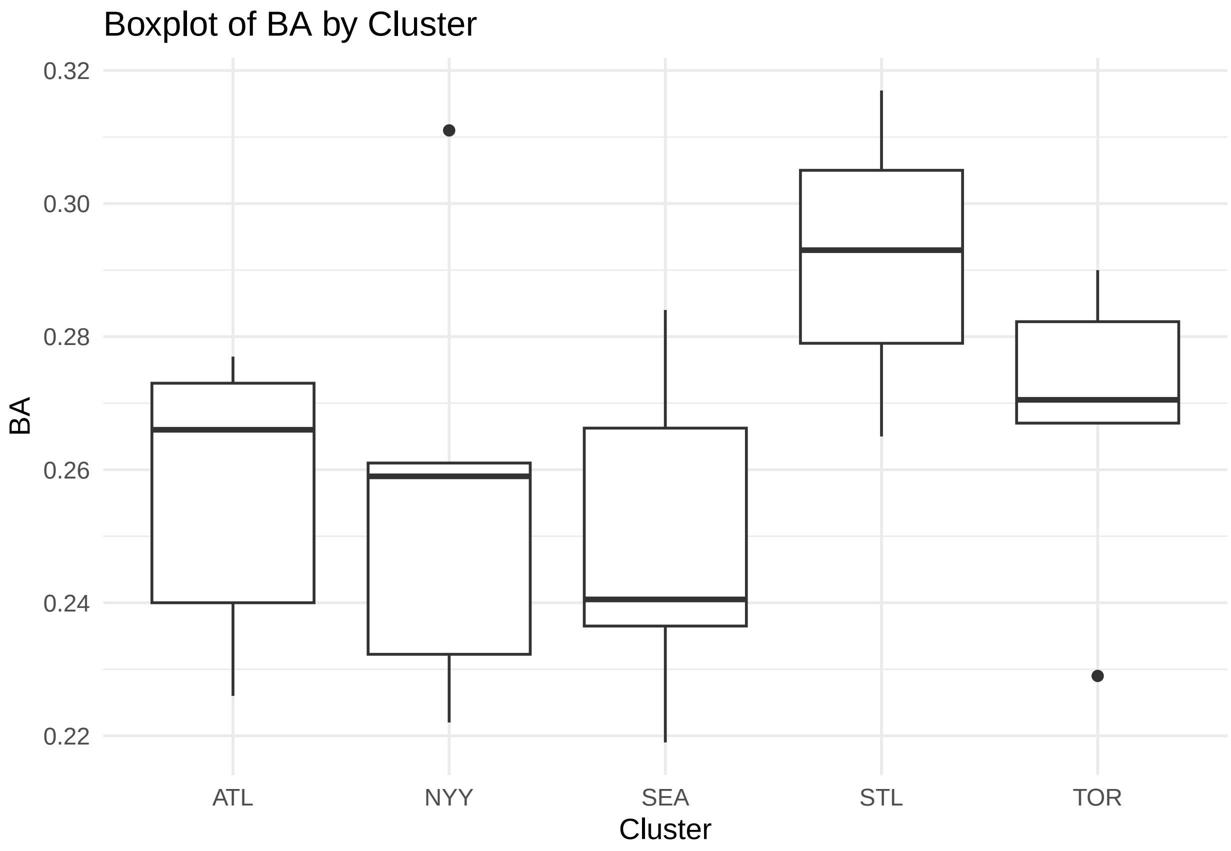
This dataset has a population size of 131, and it includes individual player information and BA data. We have confirmed through the histogram and boxplot for BA that the graph is symmetrical and there are no outliers. Since the dataset size is not large, we will calculate the estimate mean with bound using one-step cluster sampling.

```
set.seed(123)
sample(1:31, 5) # select n = 5
```

```
## [1] 31 15 19 14  3
```

```
data <- mlb %>%
  filter(Tm %in% c('STL','TOR','SEA','ATL','NYY'))

ggplot(data = data, aes(x = Tm, y = BA, group = Tm)) +
  geom_boxplot() +
  labs(title = "Boxplot of BA by Cluster", x = "Cluster", y = "BA") +
  theme_minimal()
```



```
data$Tm <- as.numeric(as.factor(data$Tm))

data <- data %>%
  group_by(Tm) %>%
  summarize(mi = n(), yi = sum(BA))
```

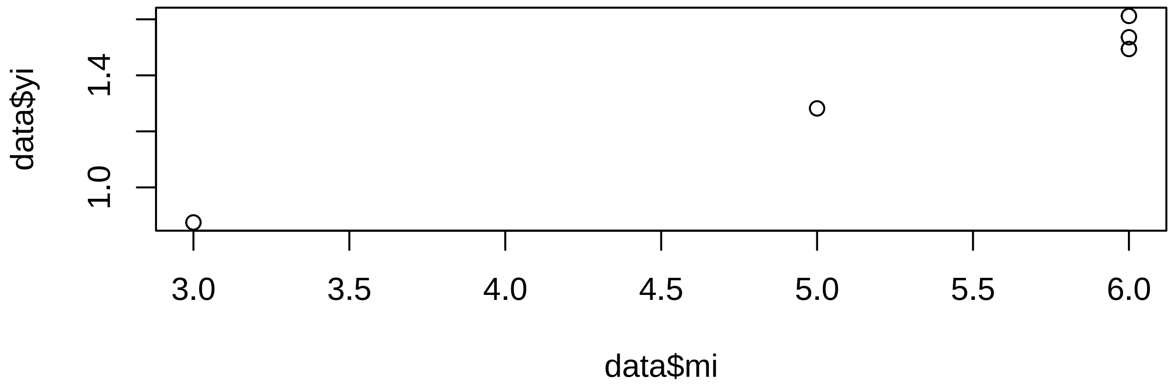
```
head(data)
```

```
## # A tibble: 5 x 3
##      Tm    mi    yi
##   <dbl> <int> <dbl>
## 1     1     5 1.28
## 2     2     6 1.54
## 3     3     6 1.49
## 4     4     3 0.875
## 5     5     6 1.61
```

## One-Stage Cluster Sampling

When the entire dataset is divided into clusters based on teams, 31 clusters are formed. We conducted sampling using Simple Random Sampling (SRS) on 5 clusters and transformed the data into a more convenient form for calculation.

```
plot(data$mi, data$yi)
```



assumption 1: Linear positive relationship between yi and mi with intercept near 0 -> Since scatterplot has positive linear trend, it satisfies assumption.

assumption 2: Component n >= 5 -> n = 5

Additionally, we have confirmed that our sample data also has no outliers and is approximately symmetric.

```
M = 131
N = 31
Mbar = M/N
n = 5

ybar = sum(data$yi) / sum(data$mi) # 0.262
```

```r
sr2 = var(data$yi - ybar*data$mi) # 0.00425
var.ybar = (1- n/N)*(sr2/(n*Mbar^2))
bound = 2*sqrt(var.ybar) # 0.013
```

As a result, population mean BA for MLB players in 2022 is between 0.262 +- 0.013

```r
# Solution?
mean(mlb$BA) # 0.258
```

```
## [1] 0.2582769
```

## Check Result

Our population mean is 0.258, and since it falls within our predicted range, we can conclude that the analysis has been performed well.

## Code Appendix

```
library(readxl)
library(dplyr)
library(tidyr)
library(ggplot2)

mlb <-read_excel("~/Desktop/JH/PSU/23 SPR/Stat 466/final project /2022_MLB_Data.xlsx")
mlb <-mlb[-94, c('Name', 'Age', 'Tm', 'BA')]
head(mlb)

par(mfrow=c(1,2))
hist(mlb$BA)
boxplot(mlb$BA)
set.seed(123)
sample(1:31, 5) # select n = 5

data <- mlb %>%
  filter(Tm %in% c('STL','TOR','SEA','ATL','NYY'))

ggplot(data = data, aes(x = Tm, y = BA, group = Tm)) +
  geom_boxplot() +
  labs(title = "Boxplot of BA by Cluster", x = "Cluster", y = "BA") +
  theme_minimal()
data$Tm <- as.numeric(as.factor(data$Tm))

data <- data %>%
  group_by(Tm) %>%
  summarize(mi = n(), yi = sum(BA))

head(data)
plot(data$mi, data$yi)
M = 131
N = 31
Mbar = M/N
n = 5

ybar = sum(data$yi) / sum(data$mi) # 0.262

sr2 = var(data$yi - ybar*data$mi) # 0.00425
var.ybar = (1- n/N)*(sr2/(n*Mbar^2))
bound = 2*sqrt(var.ybar) # 0.013
# Solution?
mean(mlb$BA) # 0.258
```