Statistical Methods – COSC 6323 - HomeWork-2

By Dinesh Narlakanti (2083649)

INTRODUCTION

More than 35,000 medals have been awarded at the Olympics since 1896. The data we are observing in this document has information about all the athletes who won medals from Athens 1896 to Rio 2016. Current data set containes various attributes like ID(unique number for each athlete), Information of the athlete like Name sex, age, height(in cms), weight(in cms) and other information like team, NOC(National Olympic Committee), Year and season, City in which the Olympic wass held, Sport and event and medal earned.

Data set is available on Kaggle and the link to the data set is: https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results

Actual data set included information about all the athletes who participated but as we require information of the athletes who earned medals only, so, deleted the rows of the athletes who didn't earn any medals at the Olympic event. Did this transformation in R by removing all the cells of the medals columns that has NA.

Modified the actual data set by assigning points to the players based on the medal they earned. Later filtered only US Swimmers from the data set by keeping NOC as United States and Sport as Swimming.

This report has information about the calculation of sample size and the variables that are used for calculation, shape of the samples' mean distribution, how the effect size affect the result and the inference from the obtained information and plotted graphs.

GETTING STARTED WITH THE HOMEWORK

Step-1: Reading the data, filtering NA, assigning the points and filtered out non-required info other than US Swimmers.

```r
athlete_data <- read.csv("C:/Users/ndine/Downloads/athelete_events.csv")

athlete_data <- athlete_data[!is.na(athlete_data$Medal),]

athlete_data$Points <- ifelse(athlete_data$Medal == 'Gold', 3,
                              ifelse( athlete_data$Medal == 'Silver', 2,
                                      ifelse( athlete_data$Medal == 'Bronze',1,0)))

USA_Swimming_Data <- subset(athlete_data, athlete_data$NOC =="USA"
                            & athlete_data$Sport =="Swimming")
```

Step-2: Now, divided the data set into two categories, Pre world war 2(before 1939) and Post world war 2(1939 and after 1939). Later aggregated the points secured as score and grouped by the names of the players.

```r
preWW <- subset(USA_Swimming_Data, USA_Swimming_Data$Year<1939)

postWW <- subset(USA_Swimming_Data, USA_Swimming_Data$Year>=1939)



pre_sum <- aggregate(preWW$Points, by=list(preWW$Name), FUN=sum)
#length(pre_sum$x)

post_sum <- aggregate(postWW$Points, by=list(postWW$Name), FUN=sum)
#length(post_sum$x)
```

Step-3: Installed pwr package and calculated sample sizes (n1 and n2) using power.t.test for medium(d=0.5) and large(0.8) effect sizes keeping power as 0.8, significance level as 0.05 and alternate hypothesis as ‘two.sided”
Effect size: It is quantitative measure of the strength of the phenomenon. As the effect size increases, the high difference between the observations are considered. d cannot be negative but there’s no limit for how large it can be.
Power: The probability of correctly rejecting a null hypothesis when it is not true. When sample size or effect size increases, power also increases but when significance level decreases, power also decreases.
Significance-level: It is closely related with confidence level. 1 - confidence level is significance level. It is the probability of wrongly rejecting the null hypothesis even when if it is true.

```
#install.packages("pwr")

p1 <- power.t.test(d=0.5,power=0.8,sig.level=0.05,alternative = "two.sided")
n1 <- p1$n

p2 <- power.t.test(d=0.8,power=0.8,sig.level=0.05,alternative = "two.sided")
n2 <- p2$n
```

Step-4: Created empty vectors to store sample means of the means.

```
sample_mean_pre_m = vector()
sample_mean_post_l = vector()
sample_mean_pre_l = vector()
sample_mean_post_m = vector()
```

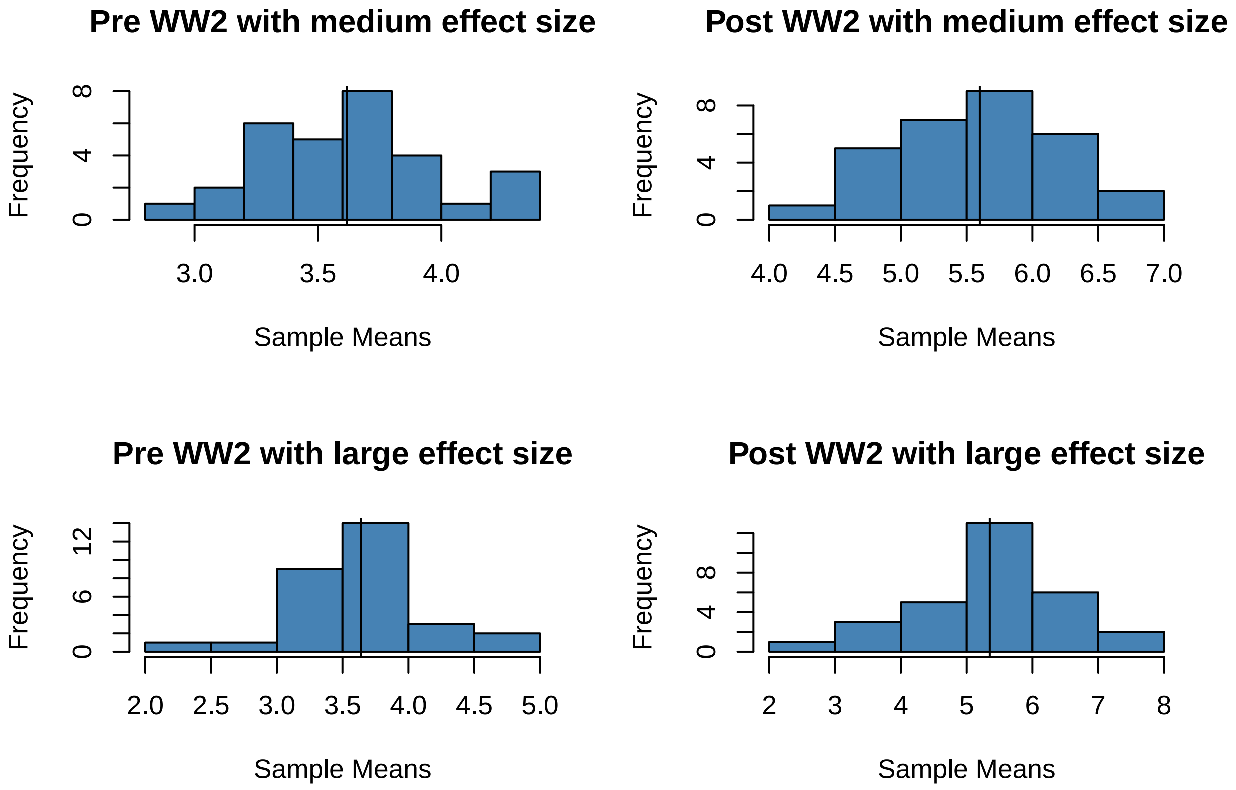
Step-5: Filled the empty vectors with mean of the sample means. Each sample size for medium is rounded to 64 and for large is rounded to 26. Values are obtained from the output of power.t.test()

```
for (i in 1:30){
  sample_mean_pre_m[i] = mean(rnorm(round(n1),mean = mean(pre_sum$x), sd=sd(pre_sum$x)))

}
for (i in 1:30){

  sample_mean_post_m[i] = mean(rnorm(round(n1),mean = mean(post_sum$x), sd=sd(post_sum$x)))
}



for (i in 1:30){
  sample_mean_pre_l[i] = mean(rnorm(round(n2),mean = mean(pre_sum$x), sd=sd(pre_sum$x)))

  }
for (i in 1:30){
  sample_mean_post_l[i] = mean(rnorm(round(n2),mean = mean(post_sum$x), sd=sd(post_sum$x)))
}
```

Step-6: Plotting the histogram of the sample mean vector

```
par(mfrow=c(2,2))
hist(sample_mean_pre_m, main = "Pre WW2 with medium effect size",
     xlab = "Sample Means", col = "steelblue")
abline(v=mean(sample_mean_pre_m))
hist(sample_mean_post_m, main = "Post WW2 with medium effect size",
     xlab = "Sample Means", col = "steelblue")
abline(v=mean(sample_mean_post_m))
hist(sample_mean_pre_l, main = "Pre WW2 with large effect size",
     xlab = "Sample Means", col = "steelblue")
```

```
abline(v=mean(sample_mean_pre_l))
hist(sample_mean_post_l, main = "Post WW2 with large effect size",
     xlab = "Sample Means", col = "steelblue")
abline(v=mean(sample_mean_post_l))
```



Question & Answers:

Q) What can you tell about the shape of the samples' means distributions?
A) As the values are taken randomly from the populated data set, the graphs that are obtained are not constant and cannot be generalized to tell the shape of the samples' mean distribution. But when i executed multiple times, most of the times i got are: Positively(Right) skewed, Unimodal, Negatively(left) skewed, Double peaked/ Bimodal, Bell-shaped.

Q)How is the effect size affecting the result (mean of the samples means)?
A) As we increase the effect size, the mean of the sample means is also increasing. Higher the effect size, higher the mean. Below are the mean of the means with medium effect size and large effect size. We can also observe it in the graph, abline() draws a vertical line at the mean.

```
mean(sample_mean_pre_m)
```

```
## [1] 3.618338
```

```
mean(sample_mean_pre_l)
```

```
## [1] 3.641279
```

```
mean(sample_mean_post_m)
```

```
## [1] 5.599499
```

```
mean(sample_mean_post_l)
```

```
## [1] 5.350516
```

Q)What can you tell about the change in the scores for two different time periods? What do you think affected the changes?
A) The change in the score was mainly because of:
i) Pre WW2 years are less than post ww2 years. As years are less, players participated and winnings are less, so that overall mean is also less in pre WW2 when compared to post WW2.

```
"Number of olypmpics conducted pre WW2"
```

```
## [1] "Number of olypmpics conducted pre WW2"
```

```
1939 - min(athlete_data$Year)
```

```
## [1] 43
```

```
"Number of olympics conducted post WW2"
```

```
## [1] "Number of olympics conducted post WW2"
```

```
max(athlete_data$Year) - 1939
```

```
## [1] 77
```

ii) Number of gold medals secured Post WW2 are more than the gold medals secured Pre WW2.

```
"Gold medals Pre WW2"
```

```
## [1] "Gold medals Pre WW2"
```

```
length(preWW$Points==3)
```

```
## [1] 137
```

```
"Gold medals post WW2"
```

```
## [1] "Gold medals post WW2"
```

```
length(postWW$Points==3)
```

```
## [1] 941
```