Statistical Methods – COSC 6323 - HomeWork-1

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 atheletes 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.

GETTING STARTED WITH THE HOMEWORK

Step-1: Started by opening a .csv file containing information with build in *read.csv(. . .)* function call, which reads in data in as a data frame and assign the data frame to a variable using <- so that it stores in R's memory. Later deleted the rows that has NA in the medal column because we need only the information of the athletes who earned medals.

```
athlete_data <- read.csv("C:/Users/ndine/Downloads/athelete_events.csv")
athlete_data <- athlete_data[!is.na(athlete_data$Medal),]
```

Step-2: Assigned points to each athlete based on the medals earned. Athlete who secured gold is assigned with 3 points, athlete with silver is 2 points and the one with bronze are given 1 point. For this, a new column to the data frame is created and name 'Points'

```
athlete_data$Points <- ifelse(athlete_data$Medal == 'Gold', 3, ifelse( athlete_data$Medal == 'Silver', 1
```

Step-3: Data is now filtered by countries and sports. Filtered out 2 countries (United States and France) with 1 sport each. 'Swimming' for USA and 'Fencing' for France. Filtering of the data is done by using 'subset'.

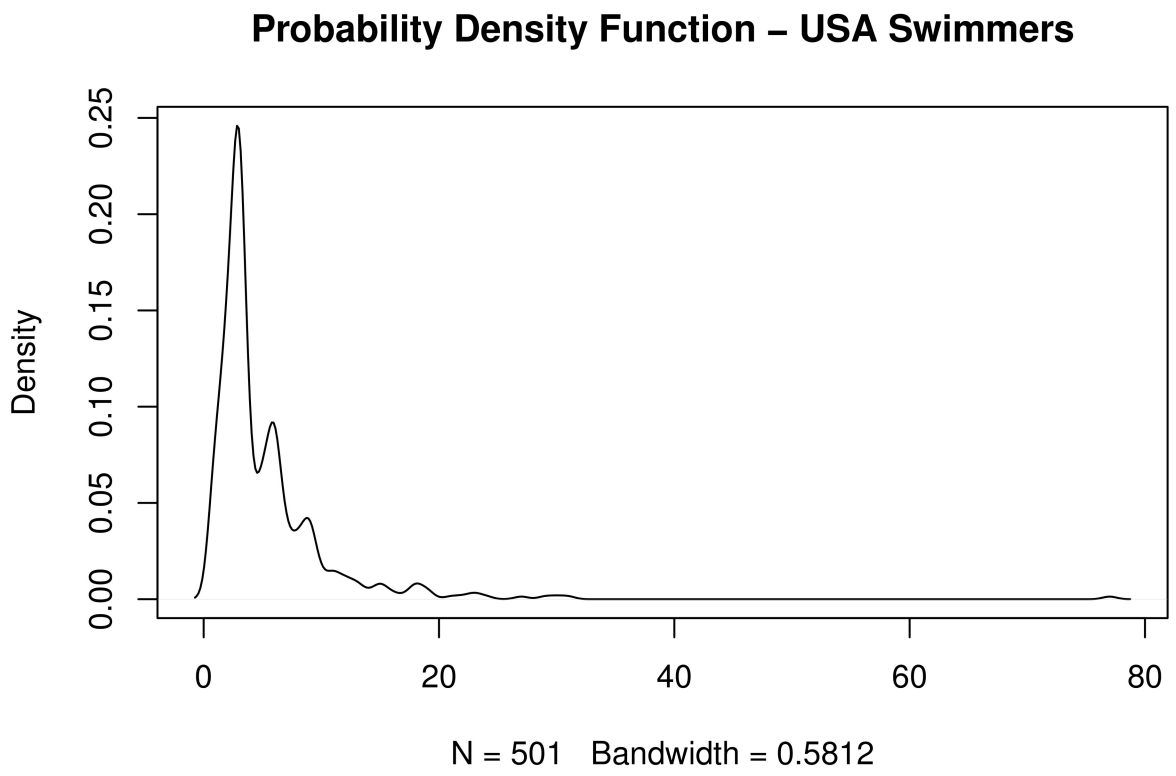
```
USA_Swimming_Data <- subset(athlete_data, athlete_data$NOC =="USA" & athlete_data$Sport =="Swimming")

France_Fencing_data <- subset(athlete_data, athlete_data$NOC =="FRA" & athlete_data$Sport =="Fencing")
```

Step-4: Aggregate functional is used to calculate the sum of the points obtained. This attribute is used while plotting the PDF and CDF.

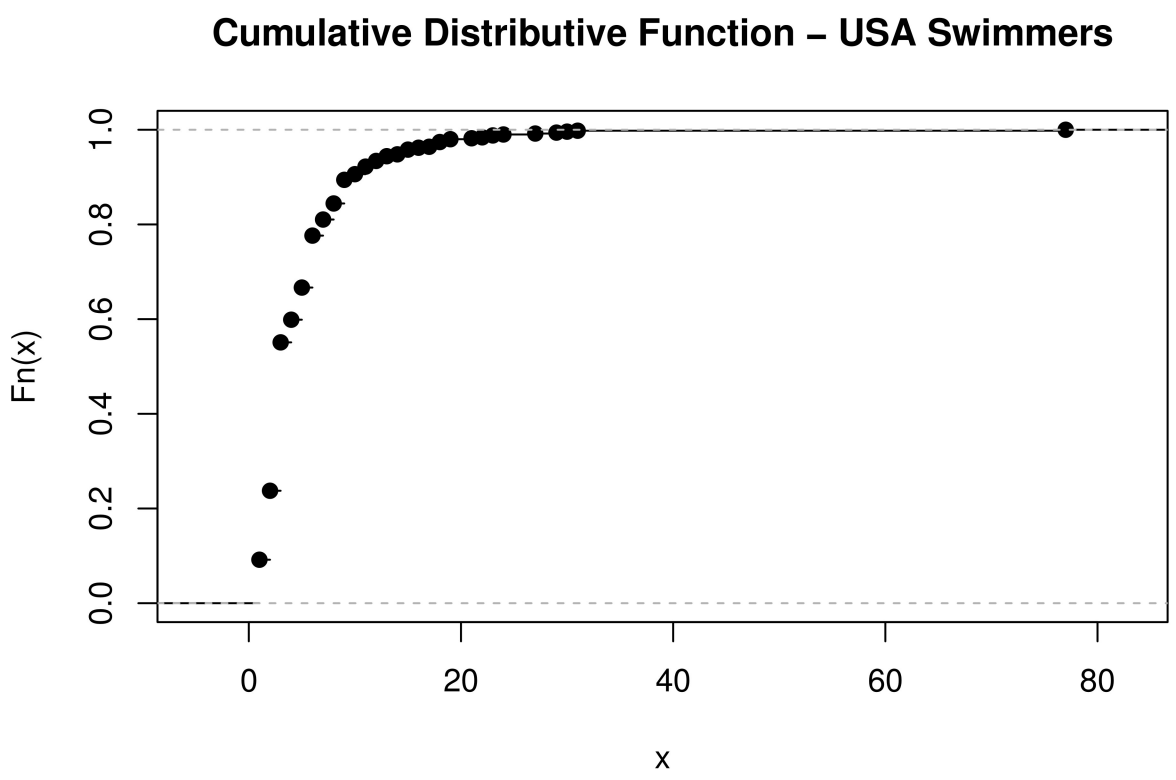
```
points_sum_usa <- aggregate(USA_Swimming_Data$Points, by=list(USA_Swimming_Data$Name), FUN=sum)

points_sum_france <- aggregate(France_Fencing_data$Points, by=list(France_Fencing_data$Name), FUN=sum)
```

Step-5: Plotted the Probability Density Function (PDF) and Cumulative Distributive Function (CDF) for the sum of points earned.

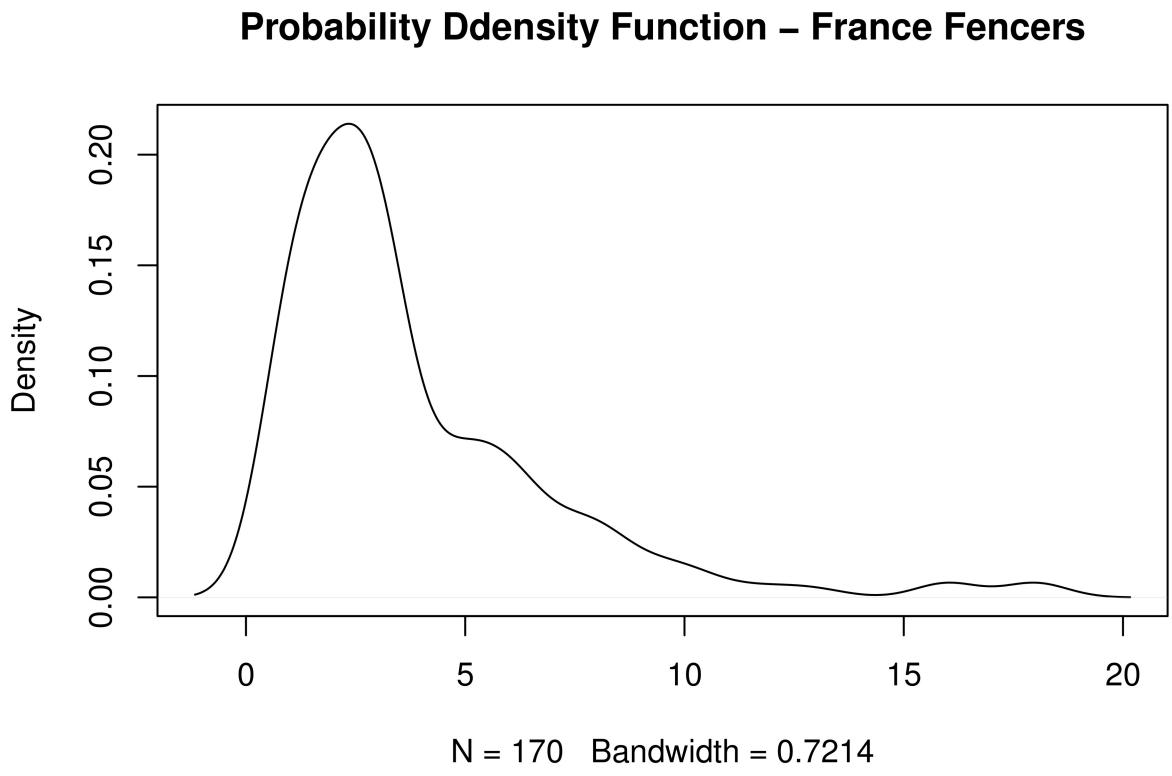
```
plot(density(points_sum_usa$x), main= "Probability Density Function - USA Swimmers")
```



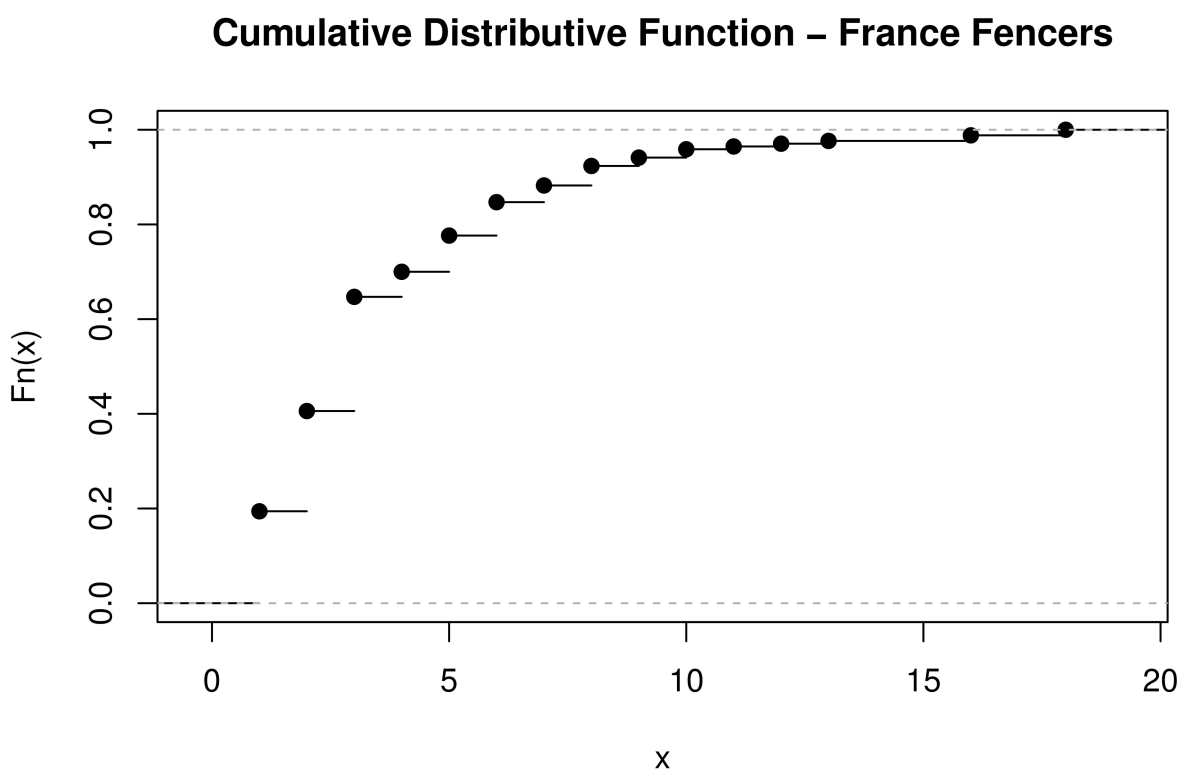

```
plot(ecdf(points_sum_usa$x), main= "Cumulative Distributive Function - USA Swimmers")
```




```
plot(density(points_sum_france$x), main= "Probability Ddensity Function - France Fencers")
```




```
plot(ecdf(points_sum_france$x), main= "Cumulative Distributive Function - France Fencers")
```



Summary of the final data set

*Summary of USA Swimmers

```
summary(points_sum_usa)
```

```
##    Group.1                x
##  Length:501         Min.   : 1.00
##  Class :character   1st Qu.: 3.00
##  Mode  :character   Median : 3.00
##                     Mean   : 5.25
##                     3rd Qu.: 6.00
##                     Max.   :77.00
```

- Summary of France Fencers

```
summary(points_sum_france)
```

```
##    Group.1                x
##  Length:170         Min.   : 1.000
##  Class :character   1st Qu.: 2.000
##  Mode  :character   Median : 3.000
##                     Mean   : 3.882
##                     3rd Qu.: 5.000
##                     Max.   :18.000
```

OBSERVATIONS & CONCLUSIONS

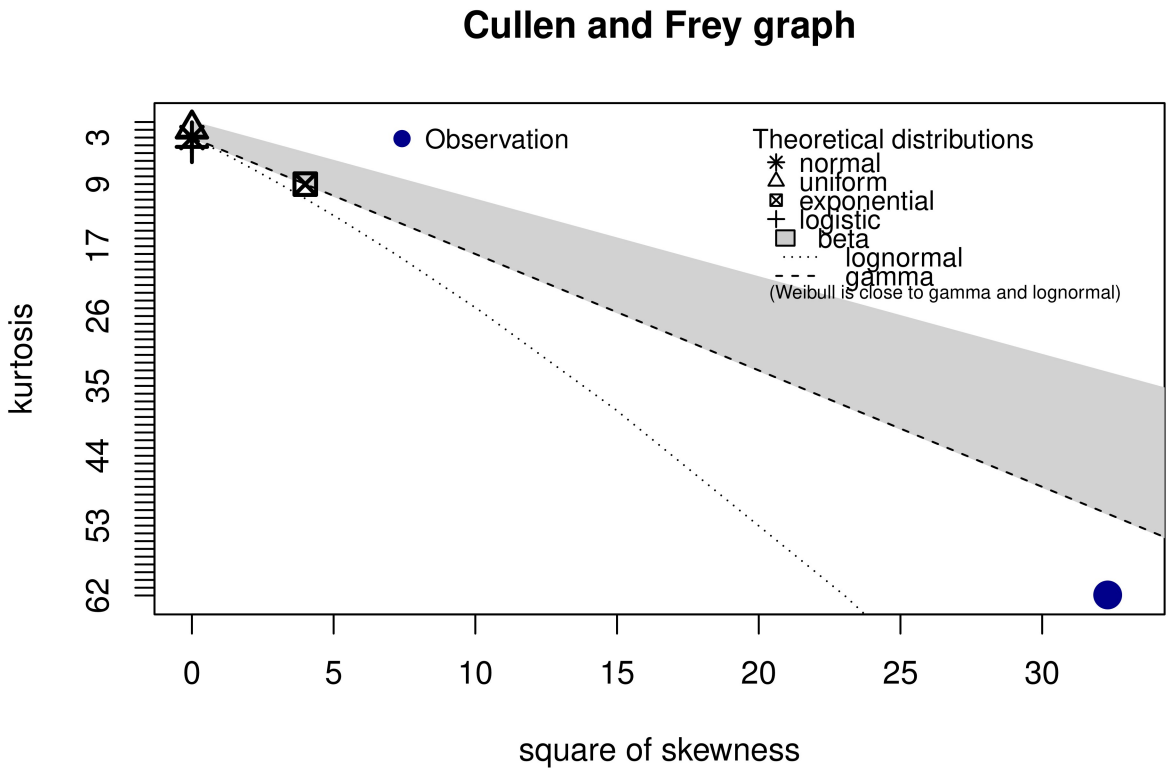
- Graph of USA Swimmers and France Fencers is resembling the graph of log normal. So, type of distribution is **"Log Normal"**
- Also, When used **descdist** function on sum of points of USA Swimmers, the observation is in between **log normal** and **gamma** distribution.

```
#install.packages("fitdistrplus")
library(fitdistrplus)
```

```
## Loading required package: MASS
```

```
## Loading required package: survival
```

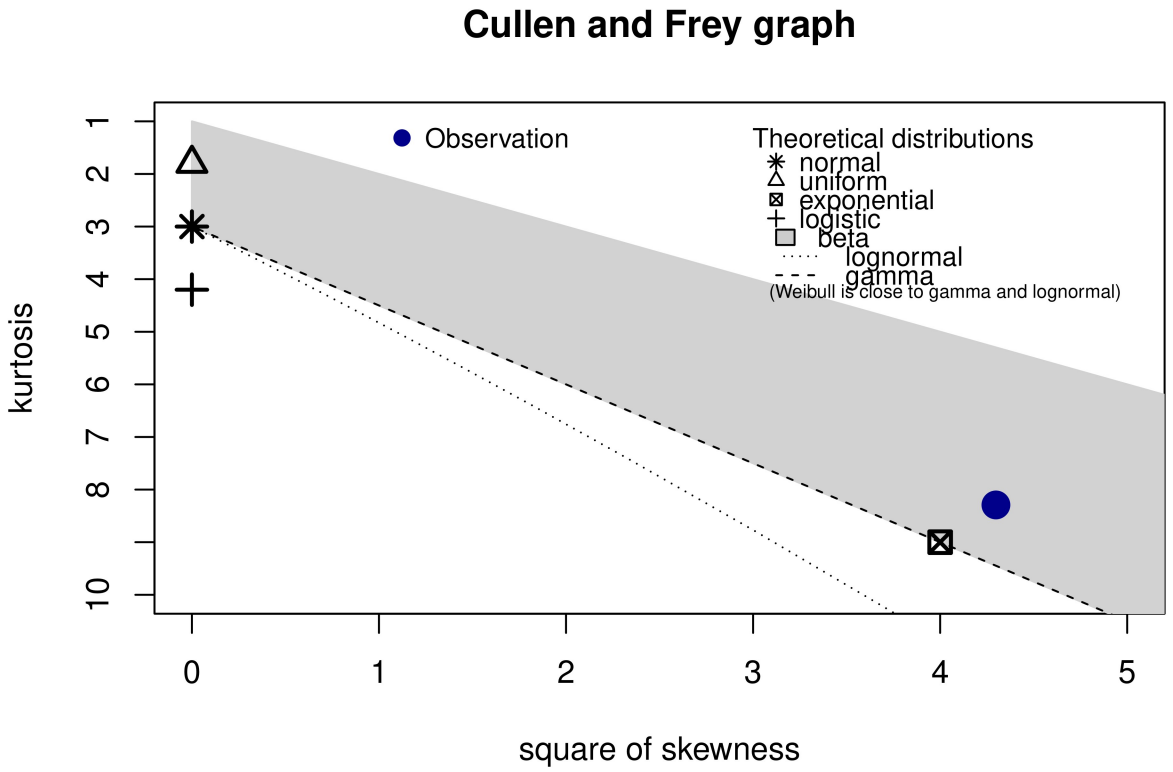
```
descdist(points_sum_usa$x, discrete = FALSE)
```




```
## summary statistics
## ------
## min:  1   max:  77
## median:  3
## mean:  5.249501
## estimated sd:  5.535668
## estimated skewness:  5.684243
## estimated kurtosis:  61.93371
```

- When used the same **descdist** function on France Fencers, the observation is clearly on **beta** distribution range.

```
descdist(points_sum_france$x, discrete = FALSE)
```




```
## summary statistics
## ------
## min:  1   max:  18
## median:  3
## mean:  3.882353
## estimated sd:  3.230446
## estimated skewness:  2.073186
## estimated kurtosis:  8.291674
```

- As the graphs we got are right(positive) skewed data, to normalise the graph, we have three methods:

1. Root: Can be transformed with high order root (Weakest transformation)
2. Logarithm: Can be transformed with the root of algorithm (Commonly used transformation)
3. Reciprocal: Can be transformed using higher exponents. (Strongest tranformation)

- While calculating the mode for sum of points obtained by both the teams, the output is 3. So, when observed the graph of PDF for both the teams, the peak is at 3.

```
v <- points_sum_usa$x
getmode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}
getmode(v)
```

```
## [1] 3
```

```
v <- points_sum_france$x
getmode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}
getmode(v)
```

```
## [1] 3
```

- From CDF of US Swimmers, we can observe and conclude that the top 20 percentile of the US Swimmers have obtained 7 or more points. We can also observe that majority of them have won 1-5 points. We can also cross check this by knowing the unique value of the sum of points and sorting them.

```
unique(points_sum_usa$x)
```

```
##  [1] 19  5  3  2  1  6 18  7 15  9 12  8 11 13  4 10 24 23 14 31 17 30 16 29 77
## [26] 22 27 21
```

```
sort(unique(points_sum_usa$x))
```

```
##  [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 21 22 23 24 27 29
## [26] 30 31 77
```

- From CDF of France Fencers, we can observe and conclude that majority of the athletes have only 1-2 points and only top 40 percentile of the France fencers have obtained 3 or more points.
