Data is the sword of the 21st century, those who wield it well, the Samurai - Jonathan Rosenberg

Statistical Analysis of Cricket Players Batting

Ishpreet Narang
Saketh Vommina Venkata
Keval Shah

TABLE OF CONTENTS

1. Abstract

The project aims at analyzing various statistics of the cricket batsmen considering various data such as scores of 50’s, 100’s, batting average, number of ducks, number of not outs, number of matches played, batting strike rate, innings played etc. Analytics would generate insights such as player consistency, playing intensity, consistency of batting success, consistency of batting average, conventional average etc. We intend to use the population data to do Hypothesis testing as well as Analysis of Variance(ANOVA)

This analysis is done for across all the generations, i.e. for every batsman to have ever played. We also aim to establish who is the “Best Batsmen of All Time” amongst all the batsmen. The tools that we would be using to implement the project include R, Excel and Minitab. We would be using R studio as our IDE, Excel for reading and writing the datasets, Minitab for understanding plots.

2. Introduction

Statistics is the tool used to derive the information from a huge data to predict, analyze and understand the trends of a field to which the data is related. It helps us use proper techniques and methods to analyze the data, engage the right analyses and efficiently present the conclusions. It helps us get a better understanding of the field. Statistics helps us analyze results based on quantitative proofs and clarify the right solutions from the dubious ones. Statistics accounts for the uncertainties and error in the results producing reliable data, analyzing the data properly and drawing reasonable conclusions.

Criclytics, the amalgam of the data analytics & cricket proves to be an essential tool to the teams to make accurate decisions in an unpredictable game such as cricket. International cricket council has 105member countries consisting close to 5,31,253 cricket players. All the players generate huge amount of data every day for 365 days in close to 5,40,290 cricket matches at 11,960 cricket grounds across the world. Circlytics helps the teams to keep an account of the players’ performances, intake of calories, training levels, interaction with fans, and much more in the pursuit of improved performance on the pitch. Circlytics provides a greater insight to the broadcasters, players & fans with more than enough background information to make a sound decision about the team’s performance.

3. Project Set-Up

3.1 Data Collection

Data for statistical analysis on batsmen from various teams of International Cricket was collected from the renowned website http://www.kaggle.com. We collected the data set regarding the desired topic to understand how batsmen are in form and how rankings are decided in ICC (International Cricket Council) for International batsmen till 2018

From the website mentioned above, we received various records of various batsmen from around the world, to be precise we used records of 1954 players who have been able to bat in One Day International Cricket format.

3.2 Data Filtering

Firstly, we started cleaning the data set by discarding the players who had an experience of a single year yielded "0" in the above new column appended. Hence, we re-filtered the players having an experience of above 3 years in the ODI format to have records which are not equal to 0.

Command use

```
new_batting_data <- filter (new_batting_data,
new_batting_data$experience != 0)
```

```
#all batsmen having 3 years or more experience
```

```
experienced_batsmen <- new_batting_data[new_batting_data$experience > 3, ]
```

In the data set which we fetched, it had many batsmen who have not yet batted in the ODI matches played by their respective international team. Hence, they had "null" records displayed under the columns of "Total Runs Scored", "number of 50's scored", "number of 100's scored", "ducks", etc.

As we used these records to implement the formulas to find "player intensity" and "consistency of batting success", the records for the batsmen who have never batted were shown to be undefined and hence those records were not required for our further analysis. So, we filtered out those records and discarded them. With that we also had to discard the players from the data set which had negative records in "consistency of batting success" as those records were not needed for the statistical analysis which we intended to do. And after sorting the batsmen list according to highest consistency of batting success to the lowest, we got down to 385 final list of players to implement our analysis on.

Command used

```
#remove everyone with negative batting success
```

```
experienced_batsmen_with_positive_batting_success <- filter(experienced_batsmen,
experienced_batsmen$batting_sucess > 0)
```

3.3 Data Creation

Then we created new records by discarding the records which were not needed or were redundant. For example, we created a new column all together known as "experience" by subtracting "career start year" from "career end year" and discarded the respective columns.

Command used

```
Experience <- new_batting_data$Career.End - new_batting_data$Career.Start
```

The next new subset we created was "playing consistency", which yielded the number of matches played over the years recorded in the experience column.

Command used

```
Playing_Intensity<-experienced_batsmen$Matches.Played/experienced_batsmen$experience
```

We used the records of playing intensity and created a new column named as "consistency of batting average". This yielded sum of number of 10s scored, 50s scored subtracting number of times the player got out on 0 and multiplying this to the playing intensity which we created.

Command used

batting_success <- (experienced_batsmen$Hundreds.Scored + experienced_batsmen$Scores.Of.Fifty.Or.More - experienced_batsmen$Ducks.Scored) * experienced_batsmen$playing_intensity

Next subset which we created was "consistency average". This column yielded the batsmen's conventional average multiplied by the playing intensity of the player respectively.

Command used

consistency_avg <- (experienced_batsmen_with_positive_batting_success$Batting.Avg * experienced_batsmen_with_positive_batting_success$playing_intensity)

4. Data Analysis

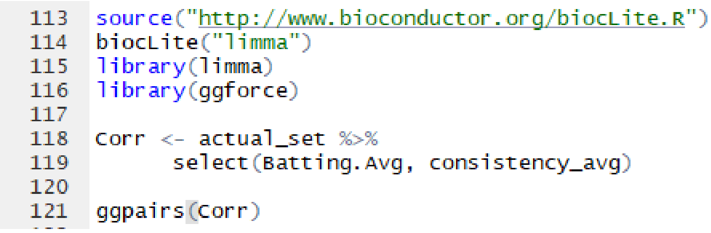
4.1 Correlation Between Traditional Average & Consistency Average

Correlation is a statistical tool to analyze to what extent the two or more variables under observation fluctuate together. There are again two types of correlations, they are as follows:

- Positive Correlation
- Negative Correlation

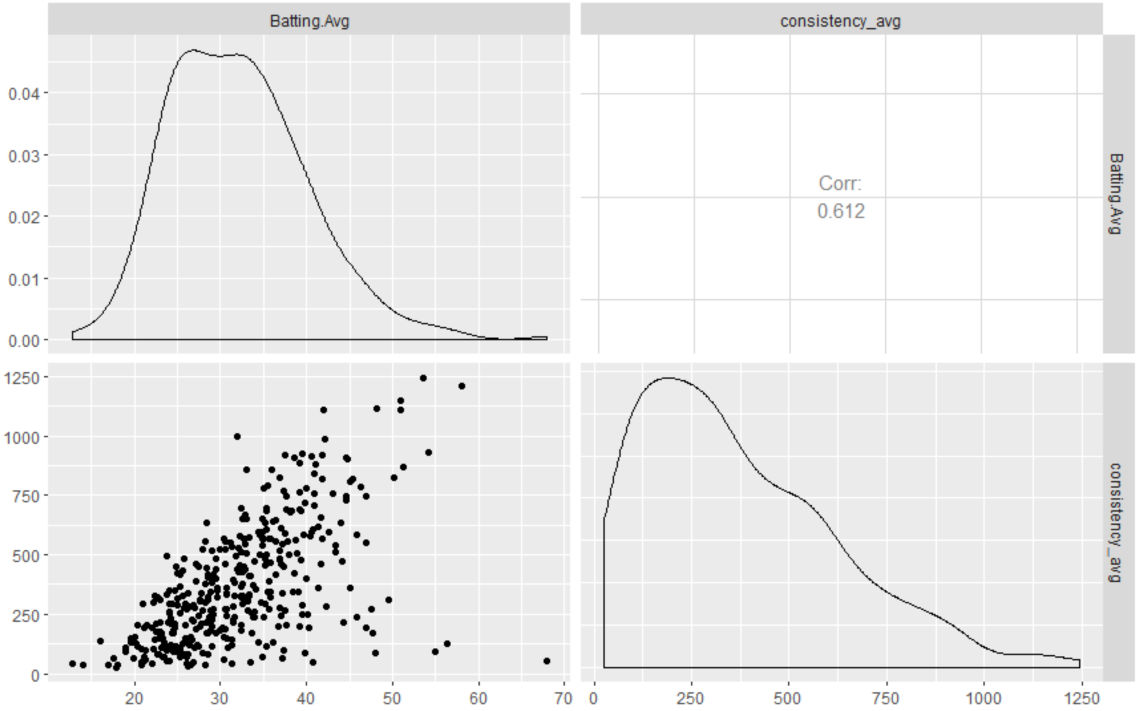
A positive correlation gives an idea about to extent to which the variables in study increase or decrease in parallel to each other. In case of a negative correlation it tells us more about the extent to which one variable increases as the other decreases and vice versa.

The variables we are interested in are traditional average and the consistency of average. We have conducted a study on how the traditional average affects the consistency average. Below is the snippet of the code that we used to find the correlation between the traditional average & the consistency average.



```
source("http://www.bioconductor.org/biocLite.R")
biocLite("limma")
library(limma)
library(ggforce)

Corr <- actual_set %>%
      select(Batting.Avg, consistency_avg)

ggpairs(Corr)
```

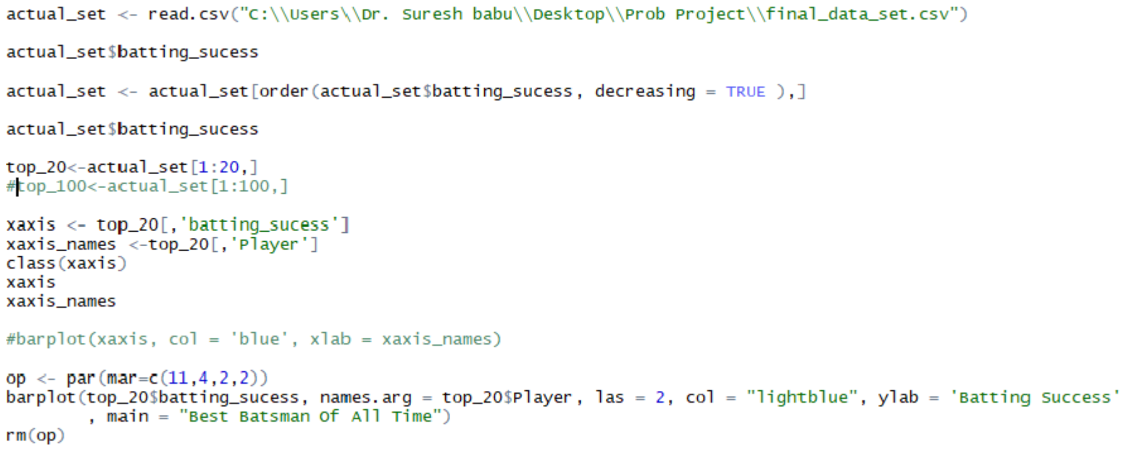
As one can see above we made use of the tool "library(ggforce)" which helps us get the visualization of the data than the statistical part of it. Below are the different data plots that we got for the batting average & the consistency average. As one can see we got a correlation of 0.612 for the variables which interprets a linear relationship between the variables somewhere between the moderate & strong.



One thing that we should keep in mind while studying the correlation is that a correlation coefficient is a statistical measure of the degree to which the changes to the value of one variable predicts the change to the value of other. It is sometimes mistaken as the causation for the change which is a wrong interpretation of it. Correlation does not imply causation.

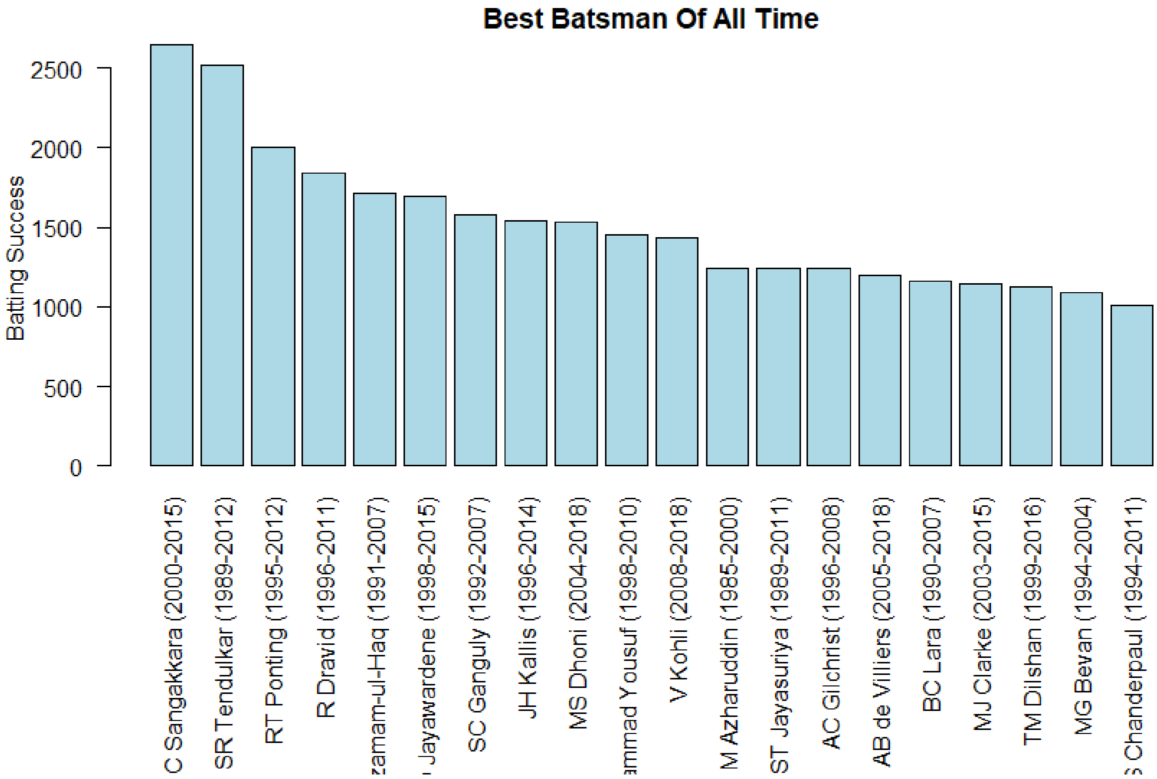
4.2 Best Batsman of All Time

The best batsman of all time was decided on the players consistency of the batting success. The consistency of batting success is defined as multiplication of the playing intensity with the addition of the no. of centuries scored to the no. of half centuries scores & subtraction of the zero score. Again, playing intensity is defined as the ratio of the total number of matches a player played to the total no. of years the player had been playing in the field. Below is a snippet of the code used to determine the above.



```
actual_set <- read.csv("C:\\Users\\Dr. Suresh babu\\Desktop\\Prob Project\\final_data_set.csv")

actual_set$batting_sucess

actual_set <- actual_set[order(actual_set$batting_sucess, decreasing = TRUE ),]

actual_set$batting_sucess

top_20<-actual_set[1:20,]
#top_100<-actual_set[1:100,]

xaxis <- top_20[,'batting_sucess']
xaxis_names <-top_20[,'Player']
class(xaxis)
xaxis
xaxis_names

#barplot(xaxis, col = 'blue', xlab = xaxis_names)

op <- par(mar=c(11,4,2,2))
barplot(top_20$batting_sucess, names.arg = top_20$Player, las = 2, col = "lightblue", ylab = 'Batting Success'
        , main = "Best Batsman Of All Time")
rm(op)
```

It is a common perception amongst the cricketing world that the Sachin Tendulkar is the best batsman of all the time, but the data paints a different picture all together. After analyzing the data, we had it came to the light that C Sangakkara is the best player of all the time and Sachin Tendulkar is the second best. Below one can see the barplot of the players versus the batting success.

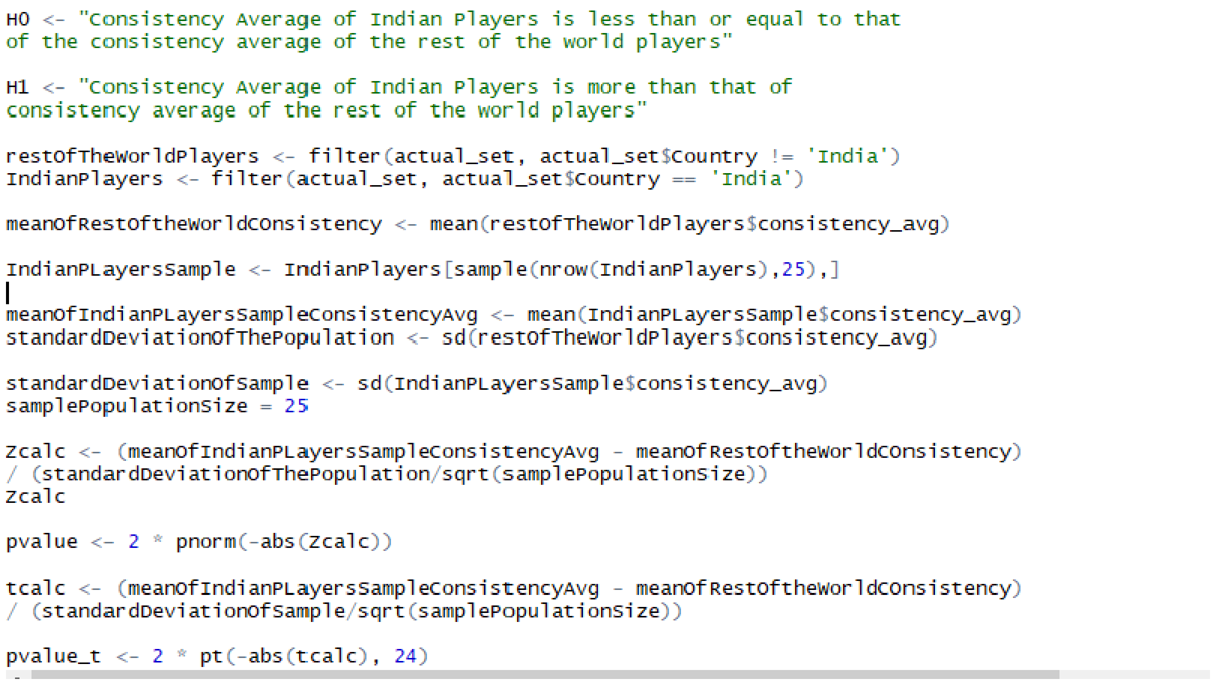


4.3 Hypothesis

Hypothesis testing is a statistical method that is used in making statistical decisions using experimental data. It was introduced by Ronald Fisher, JeryNeyman, Karl Pearson and Egon Pearson. Some of the key concepts are: -

- Null Hypothesis: - Null Hypothesis is a statistical hypothesis that the observation is due to a chance factor. It is denoted by Ho.
- Alternate Hypothesis: -It is contrary to null hypothesis and is denoted by H1. Alternative hypothesis shows that observations are the result of a real effect.
- Level of significance: -Refers to the degree of significance in which we accept or reject the null-hypothesis. 100% accuracy is not possible for accepting or rejecting a hypothesis.

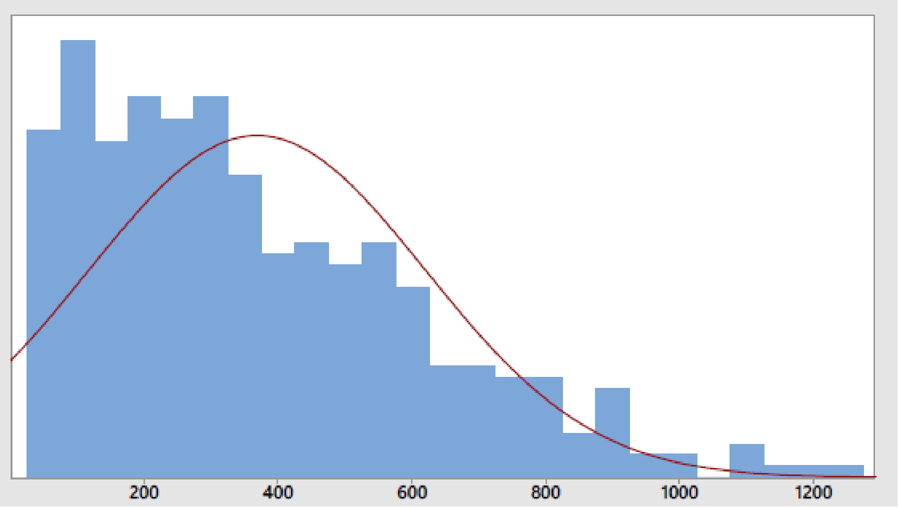
Below is the snippet of the code, where Hypothesis testing has been performed on the data set using R.



```
H0 <- "Consistency Average of Indian Players is less than or equal to that
of the consistency average of the rest of the world players"

H1 <- "Consistency Average of Indian Players is more than that of
consistency average of the rest of the world players"

restOfTheWorldPlayers <- filter(actual_set, actual_set$Country != 'India')
IndianPlayers <- filter(actual_set, actual_set$Country == 'India')

meanOfRestOftheWorldConsistency <- mean(restOfTheWorldPlayers$consistency_avg)

IndianPLayersSample <- IndianPlayers[sample(nrow(IndianPlayers),25),]

meanOfIndianPLayersSampleConsistencyAvg <- mean(IndianPLayersSample$consistency_avg)
standardDeviationOfThePopulation <- sd(restOfTheWorldPlayers$consistency_avg)

standardDeviationOfSample <- sd(IndianPLayersSample$consistency_avg)
samplePopulationSize = 25

Zcalc <- (meanOfIndianPLayersSampleConsistencyAvg - meanOfRestOftheWorldConsistency)
/ (standardDeviationOfThePopulation/sqrt(samplePopulationSize))
Zcalc

pvalue <- 2 * pnorm(-abs(Zcalc))

tcalc <- (meanOfIndianPLayersSampleConsistencyAvg - meanOfRestOftheWorldConsistency)
/ (standardDeviationOfSample/sqrt(samplePopulationSize))

pvalue_t <- 2 * pt(-abs(tcalc), 24)
```

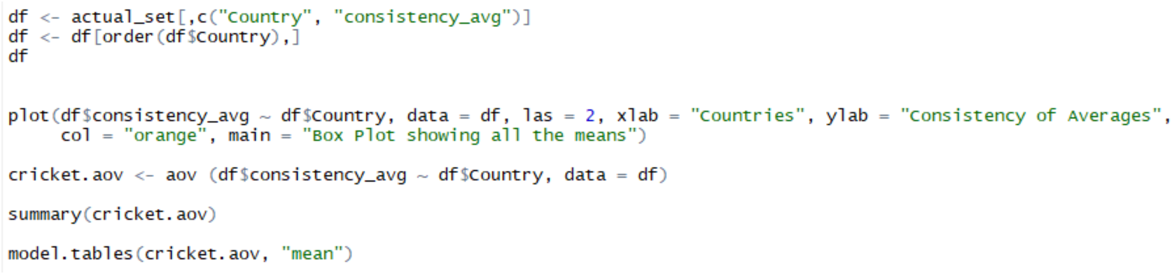
The Confidence Interval considered for the hypothesis is 95%. According to our hypothesis the test to be performed was a Right-tailed Test. P-value approach was followed to find the hypothesis. Z distribution – Standard deviation of the population was calculated from the data set and Zcalc was found. This value was found to be 0.752. The corresponding p-value was computed to be 0.451. Since p-value is greater than alpha (0.05), we **fail to reject H0**. T distribution – Here, the standard deviation of the population was unknown. We used the standard deviation of the sample to find the Tcalc value which was computed to be 0.624. The corresponding p-value was 0.538. Since p-value is greater than alpha (0.05), we **fail to reject H0.**

Therefore, from the test we conclude that "Consistency Average of Indian Players is less than or equal to that of the consistency average of the rest of the world players" with a confidence interval of 95%. The distribution is as follows:



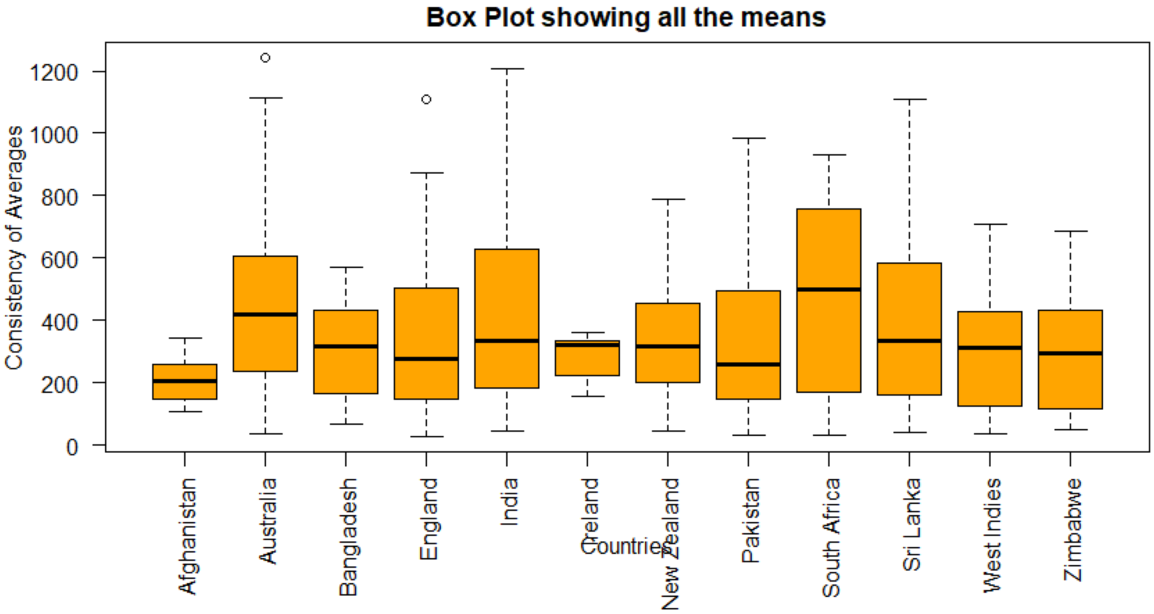
4.4 ANOVA

Analysis is a collection of statistical models and their associated estimation procedures used to analyze the differences among group means in a sample. ANOVA was developed by statistician and evolutionary biologist Ronald Fisher. In its simplest form, ANOVA provides a statistical test of whether the population means of several groups are equal, and therefore generalizes the *t*-test to more than two groups. We use One-Way ANOVA in this project. The null hypothesis for any ANOVA is that “**all population means are exactly equal”**. The following is the code snippet for finding ANOVA:

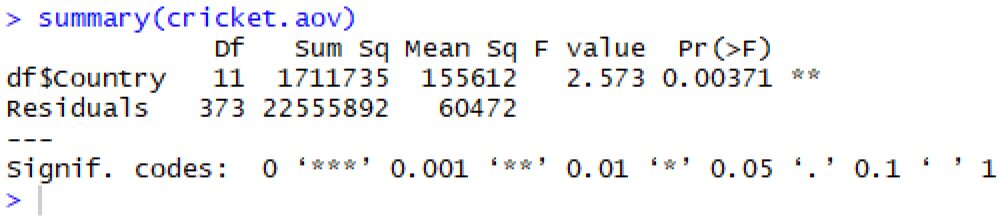


```
df <- actual_set[,c("Country", "consistency_avg")]
df <- df[order(df$Country),]
df


plot(df$consistency_avg ~ df$Country, data = df, las = 2, xlab = "Countries", ylab = "Consistency of Averages",
     col = "orange", main = "Box Plot showing all the means")

cricket.aov <- aov (df$consistency_avg ~ df$Country, data = df)

summary(cricket.aov)

model.tables(cricket.aov, "mean")
```

The box-plot of all the means is as follows:

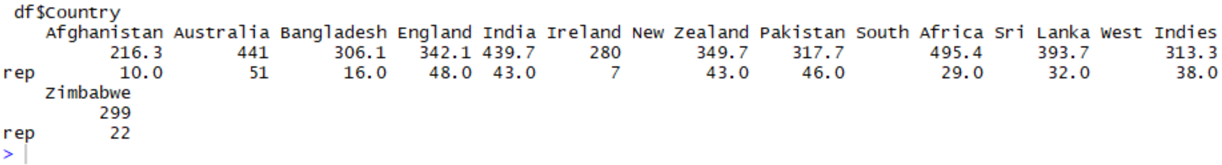


Summary of **ANOVA**:



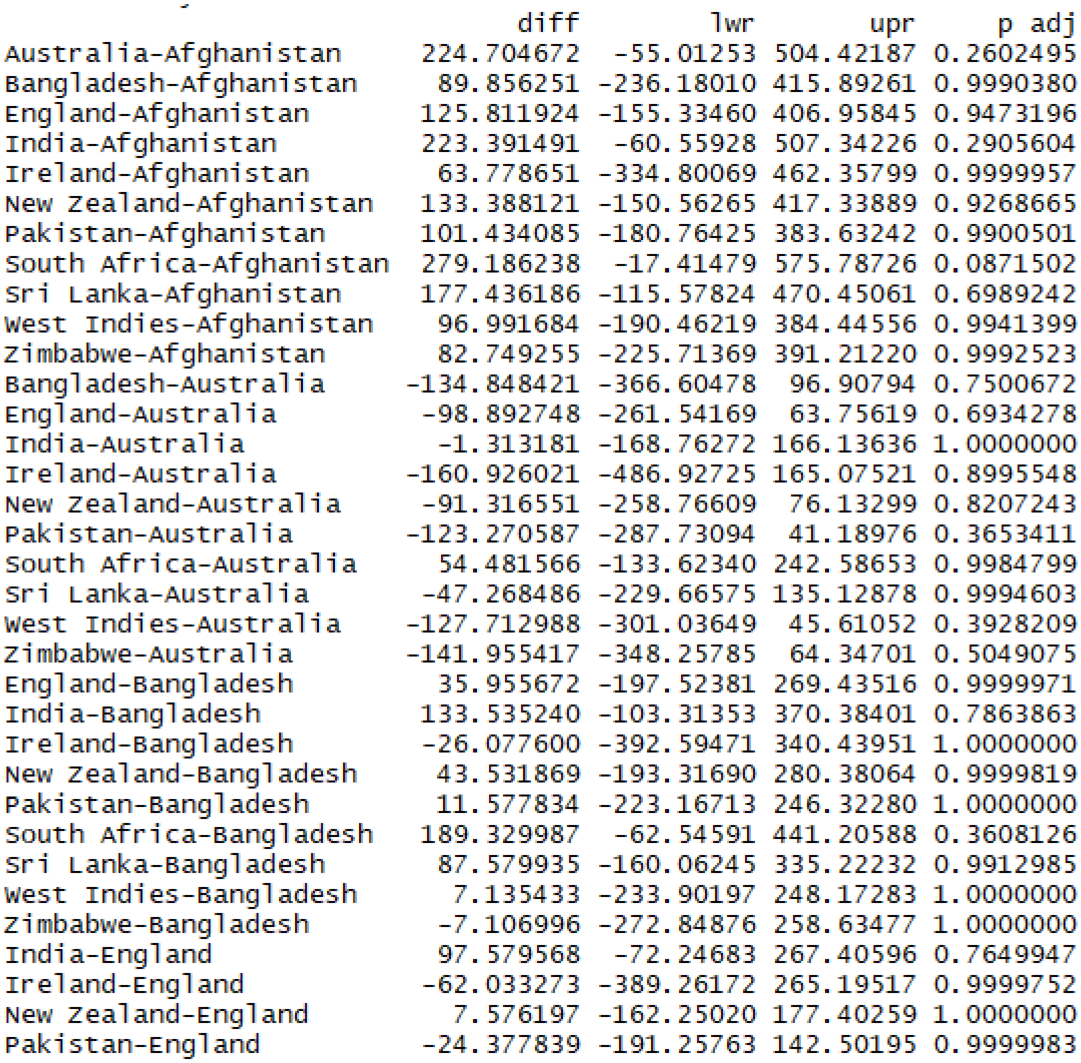
```
> summary(cricket.aov)
             Df   Sum Sq Mean Sq F value  Pr(>F)   
df$Country   11  1711735  155612   2.573 0.00371 **
Residuals   373 22555892   60472                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
```

The mean table for all countries is:



```
 df$Country
    Afghanistan Australia Bangladesh England India Ireland New Zealand Pakistan South Africa Sri Lanka West Indies
          216.3       441      306.1   342.1 439.7     280       349.7    317.7        495.4     393.7       313.3
rep        10.0        51       16.0    48.0  43.0       7        43.0     46.0         29.0      32.0        38.0
    Zimbabwe
         299
rep       22
> 
```

ANOVA clearly indicates that the means of all the samples are quite different, but we don't know how exactly they differ. To find this we do post hoc test – Tukey HSD. A sample display of the analysis is as follows:



```
                                 diff        lwr       upr     p adj
Australia-Afghanistan      224.704672  -55.01253 504.42187 0.2602495
Bangladesh-Afghanistan      89.856251 -236.18010 415.89261 0.9990380
England-Afghanistan        125.811924 -155.33460 406.95845 0.9473196
India-Afghanistan          223.391491  -60.55928 507.34226 0.2905604
Ireland-Afghanistan         63.778651 -334.80069 462.35799 0.9999957
New Zealand-Afghanistan    133.388121 -150.56265 417.33889 0.9268665
Pakistan-Afghanistan       101.434085 -180.76425 383.63242 0.9900501
South Africa-Afghanistan   279.186238  -17.41479 575.78726 0.0871502
Sri Lanka-Afghanistan      177.436186 -115.57824 470.45061 0.6989242
West Indies-Afghanistan     96.991684 -190.46219 384.44556 0.9941399
Zimbabwe-Afghanistan        82.749255 -225.71369 391.21220 0.9992523
Bangladesh-Australia      -134.848421 -366.60478  96.90794 0.7500672
England-Australia          -98.892748 -261.54169  63.75619 0.6934278
India-Australia             -1.313181 -168.76272 166.13636 1.0000000
Ireland-Australia         -160.926021 -486.92725 165.07521 0.8995548
New Zealand-Australia      -91.316551 -258.76609  76.13299 0.8207243
Pakistan-Australia        -123.270587 -287.73094  41.18976 0.3653411
South Africa-Australia      54.481566 -133.62340 242.58653 0.9984799
Sri Lanka-Australia        -47.268486 -229.66575 135.12878 0.9994603
West Indies-Australia     -127.712988 -301.03649  45.61052 0.3928209
Zimbabwe-Australia        -141.955417 -348.25785  64.34701 0.5049075
England-Bangladesh          35.955672 -197.52381 269.43516 0.9999971
India-Bangladesh           133.535240 -103.31353 370.38401 0.7863863
Ireland-Bangladesh         -26.077600 -392.59471 340.43951 1.0000000
New Zealand-Bangladesh      43.531869 -193.31690 280.38064 0.9999819
Pakistan-Bangladesh         11.577834 -223.16713 246.32280 1.0000000
South Africa-Bangladesh    189.329987  -62.54591 441.20588 0.3608126
Sri Lanka-Bangladesh        87.579935 -160.06245 335.22232 0.9912985
West Indies-Bangladesh       7.135433 -233.90197 248.17283 1.0000000
Zimbabwe-Bangladesh         -7.106996 -272.84876 258.63477 1.0000000
India-England               97.579568  -72.24683 267.40596 0.7649947
Ireland-England            -62.033273 -389.26172 265.19517 0.9999752
New Zealand-England          7.576197 -162.25020 177.40259 1.0000000
Pakistan-England           -24.377839 -191.25763 142.50195 0.9999983
```

From Tukey the following insights were uncovered:

Most teams have significant statistical difference against the Afghanistan Cricket team. Amongst the more established international cricket teams, significant statistical difference was found between South African Cricket Team and the Pakistan Cricket Team.

The two most dominant teams in the last few decades were the Indian and Australian Cricket team, which is also shown in their statistics. There is an extremely minute difference between the two teams in terms of their statistics.

Similarly, other teams with such close statistical differences are:

- Ireland – Bangladesh
- Pakistan – Bangladesh
- West Indies – Bangladesh
- Zimbabwe – Bangladesh
- New Zealand – England
- West Indies -Ireland
- Zimbabwe – Ireland
- West Indies – Pakistan
- Zimbabwe – Pakistan
- Zimbabwe - Pakistan

5. References

- http://www.simafore.com/blog/bid/78198/Best-batsman-of-all-time-fun-with-cricket-statistics
- R Lab Manuscript
- https://www.theanalysisfactor.com/r-11-bar-charts
- https://www.rdocumentation.org
- https://www.statisticssolutions.com/hypothesis-testing
- https://whatis.techtarget.com/definition/correlation
- http://statisticsbyjim.com/basics/importance-statistics
- https://www.simplilearn.com/how-big-data-is-helping-teams-win-big-at-t20-world-cup-criclytics-article
- https://www.dummies.com/education/math/statistics/how-to-interpret-a-correlation-coefficient-r/