An Analytical Study on Football Players in FIFA'19 Game: Statistical Approach

Amal Sharma

Submitted to : Shivani Patel & Savan Kumar

Course: IE 6200

Table of Contents

Introduction

FIFA’19 is a football simulation video game developed by EA Vancouver. Analysis in this report is done for the FIFA'19 data set where player statistics along with overall ranking, wage, strength, and stamina is given. Data set contains information about 18159 players which represent different countries and clubs. Football is the most popular game played in the world with viewers more than millions, which fascinated my interest in analytical study in this field.

Insight and relationships between Player's nationality, Age, Height, Penalties can be drawn from the dataset. Moreover, some interpretations can be drawn by drilling down the data. This observational study is designed to:

A. Find the pattern in Work Rate for Football players from Spain and England

B. The average age for Spanish Players playing football

C. Find Proportion between Left foot and Right foot players having real face in the FIFA’19

D. Comparison of Stamina between Football players from Spain and England

E. Find pattern between Age & penalty, Age & Stamina and Age & Strength

The goal of this observational study is to provide estimations for these questions below:

1. What is the average age for Spanish players playing football?
2. Which features are
3. highly correlated with a player's age?
4. Whether players Preferred Left foot or Right Foot while playing football?
5. Is there a difference in the Stamina between Players from England and Players from Spain?

6. Whether Work rate has a role to play in defining the dynamics of the football game?

Variable Definition:

- **Nationality**: It is the main category that represents the players' information on the country they play football from. A lot of analysis depends and can be chalked out from this variable. For eg: Which country has the maximum number of players and what is the mean age for them?
    - The type of the variable: Categories Variable; Explanatory Variable; Discrete Variable.
    - The scale of the variable: This observational study focuses on the top 10 countries with maximum players.

- **Age:** This variable is critical and defines the relationship with Strength, Stamina, and Penalty.
    - The type of the variable: Quantitative Variable; Response Variable; Discrete Variable.
    - The scale of the variable: This study has data set from 20 to 40 years of age

- **Foot Preference:** Definition: Foot is an important category in the game. Some of the Football players have more strength in the left foot whereas some have more in the right foot. This study focuses on defining the proportion in our data set
    - The type of the variable: Categories variable; Explanatory variable; Discrete variable.
    - The scale of the variable: Players has to be in two categories which are either left or right foot

- **Work Rate:** Definition: This observational study narrows Work Rate into 9 categories based on players Preference

- ○ The type of the variable: Categories variable, Explanatory variable, Discrete variable.
- ○ The scale of the variable: Participants are in any of the 9 categories which is mentioned below

Work Rate	**Count**
High/ High	34
High/ Low	18
High/ Medium	109
Low/ High	14
Low/ Low	2
Low/ Medium	8
Medium/ High	62
Medium/ Low	25
Medium/ Medium	328

- **Stamina:** Stamina has a relationship with nationality and Age. Our study focuses to identify the relationship between 2 different nationalities and does stamina changes with respect to that

  - ○ The type of the variable: Quantitative Variable; Response Variable; Discrete Variable.
  - ○ The scale of the variable: This study has 600 different data points

- **Strength:** Strength has a relationship with Age. Our study focuses on how does Strength changes with respect to age. This is really interesting data
    - The type of the variable: Quantitative Variable; Response Variable; Discrete Variable.
    - The scale of the variable: This study has 18159 different data points ranging from 17 to 97

- **Penalty:** The Main purpose of this variable is to identify the distribution of penalties among different ages. I want to analyze the relationship between them
    - The type of the variable: Quantitative Variable; Response Variable; Discrete Variable.
    - The scale of the variable: This study has 18159 different data points ranging from 5 to 92

Methods:

- The target population for this study: Football Players from FIFA'19 game station
    - Sampling Size: 18159 data points.
    - Sample Frame: Participants need to be over 18 years old, and need to be in the FIFA'19 Football game

Sampling Strategy:

- While Doing the Statistical analysis, we have defined small subsets from the main 18159 data points.

- Secondly, analysts used a sample frame to rule out data points from nationality which has fewer data points and has less potential of strong relationship through Exploratory analysis

- Thirdly, analysts used a stratified sampling strategy for statistical analysis. Stratified sampling divides potential participants into two nationality groups. It randomly selects 300 players from the Spanish group and 300 players from the English group for two nationality groups comparison.

- Lastly, we used the shuffling method to avoid biases and remove relationships

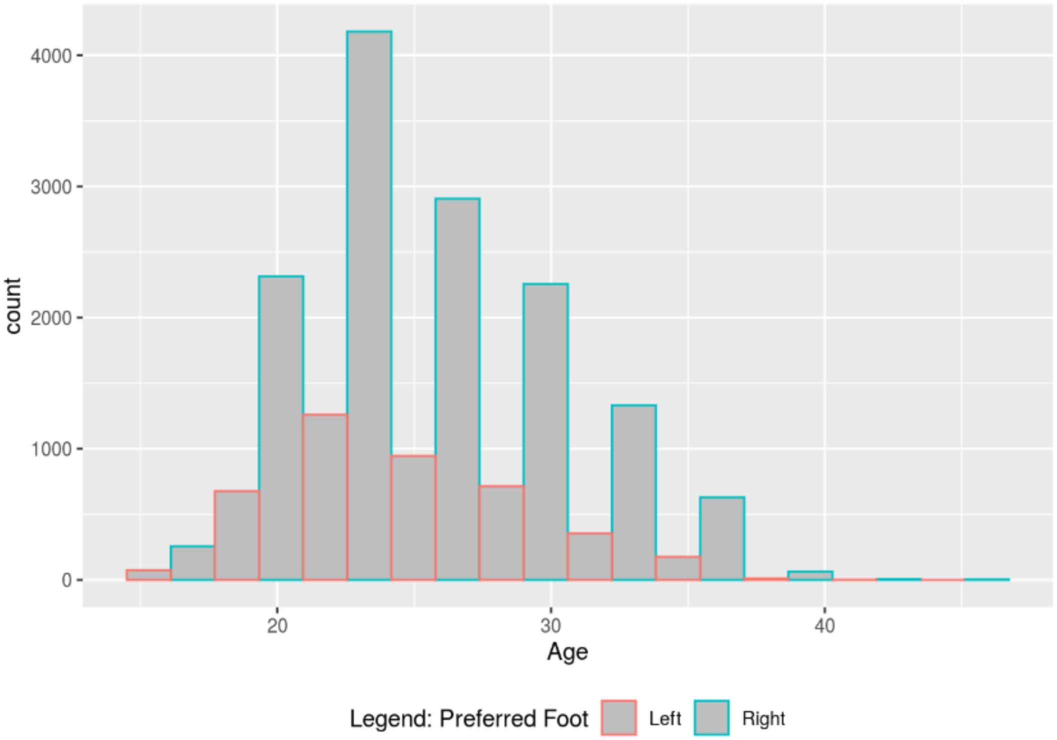
Randomization:

- Randomly selects 300 players from the Spanish team and 300 players from the England team for two nationality groups comparison.
- Randomly we transformed and shuffled data to avoid biases

The shortcoming of sample strategy:
Following Bias can be observed:
- Information Bias – Information given in the dataset is a dummy and is not related to real-world data
- Selection Bias-The sample size may not be a representation of actual Spanish or English players

Analysis

Exploratory Analysis

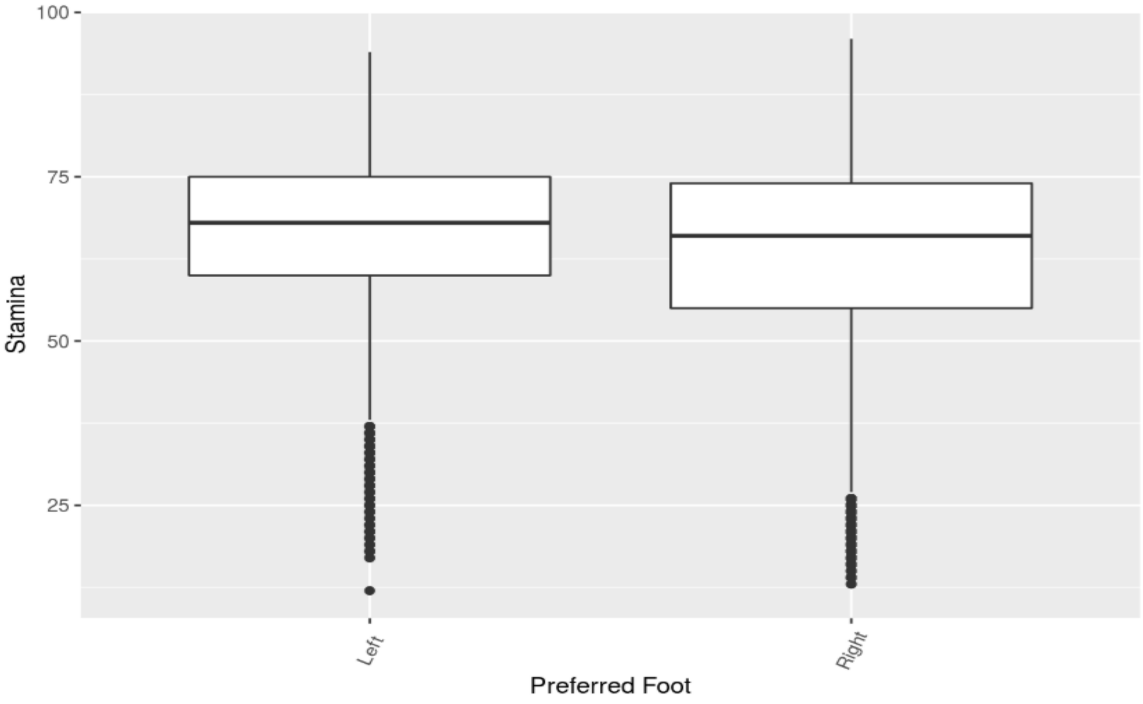


(i) Age distribution

We analyzed data to find out age distribution for the FIFA'19 football players and where is the data most populated. The graph presented above also shows distribution among left and right foot players along with age having a mean of 25.12 years and a standard deviation of 4.67. It is clearly visible that data is skewed more between 20-30 with the majority of the players uses the right foot for playing football.

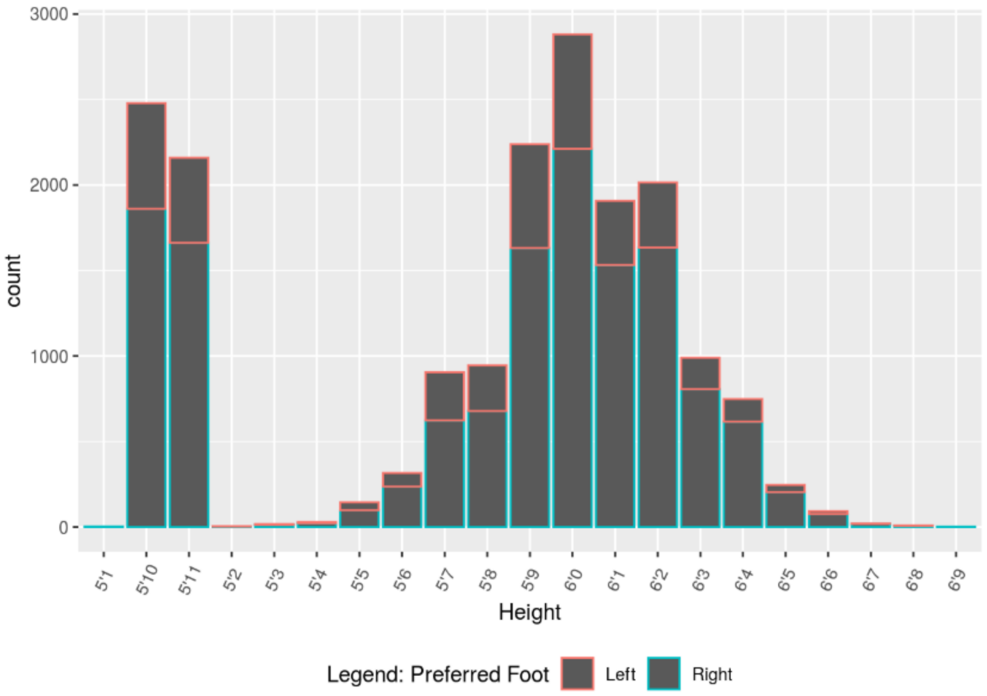
Preferred Foot V/s Stamina:

Comparing the preferred foot in terms of Stamina, we found that though the Right foot is preferred by more players, Left foot players have more stamina around 69. In addition, Left Foot players have more outliers which can have an impact on our study.



(ii) Age Group vs Population distribution

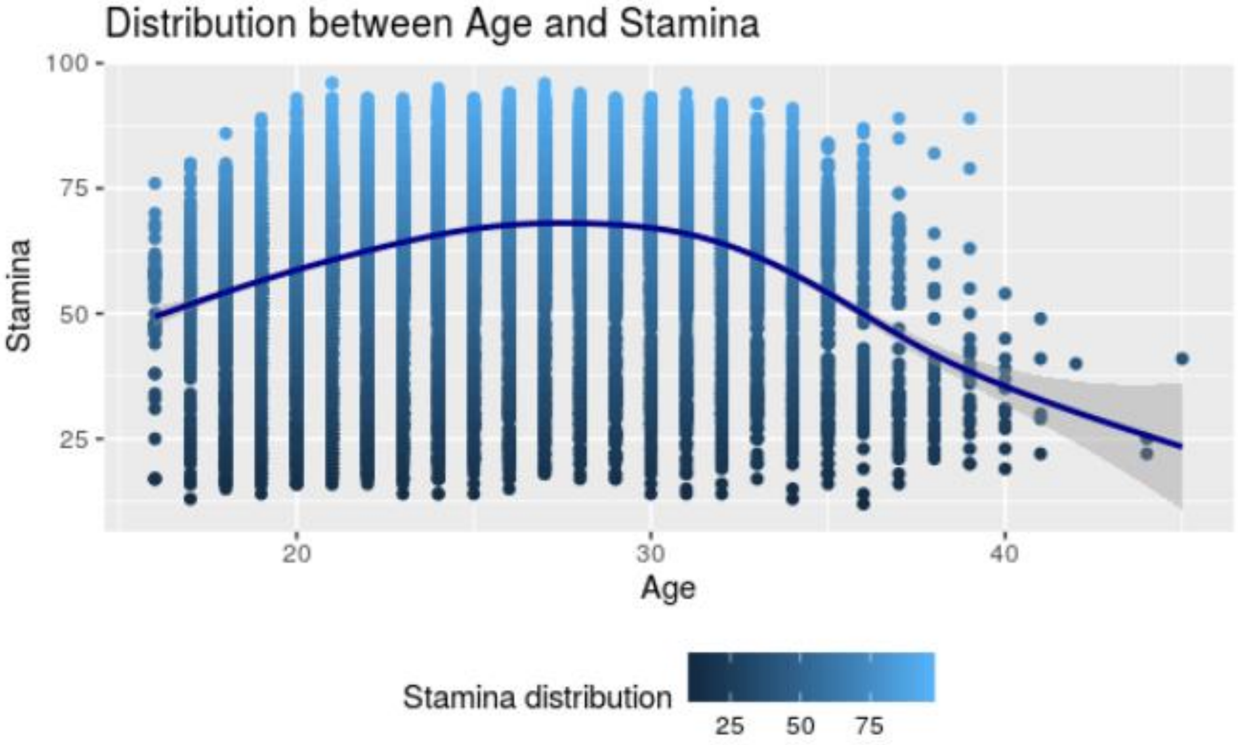
Height distribution with respect to Preferred foot:



(iii) Height Distribution wrt preferred foot

On further analysis, this study found that preferred foot has a relationship with the height of the players as well. The left foot which is quite a few in numbers has a presence in almost all of the heights. Players have the most presence around 6-foot height with almost 80% constituted by the right foot. Though there are few outliers towards the left and right side which can be ignored

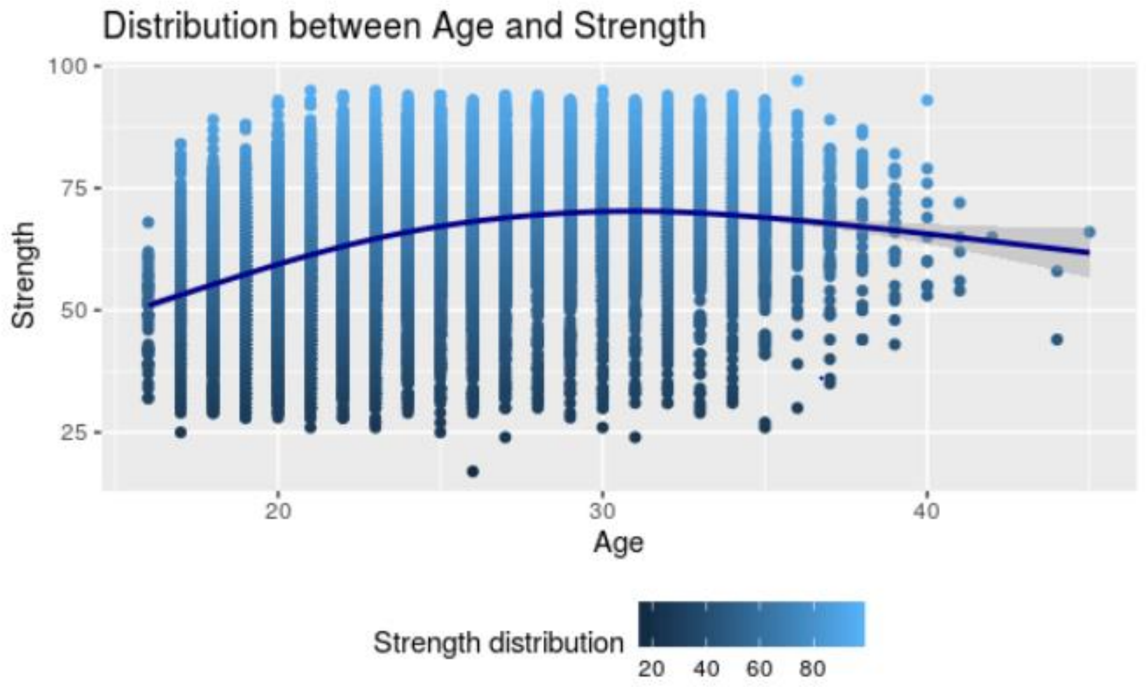
Age VS Player’s Stamina:



(iv) Age vs Stamina

It is instinctive to think that player's performance hit with age and experience and that their stamina would reflect in this relationship. The graph above shows 2 different sides of stamina with respect to age, their stamina tends to increase up to their late 20s and then begin to decline in between 30-40 years of age. It also shows that before 20s stamina was improving and reached a saturation point in mid-20s.

Age VS Player's Strength:

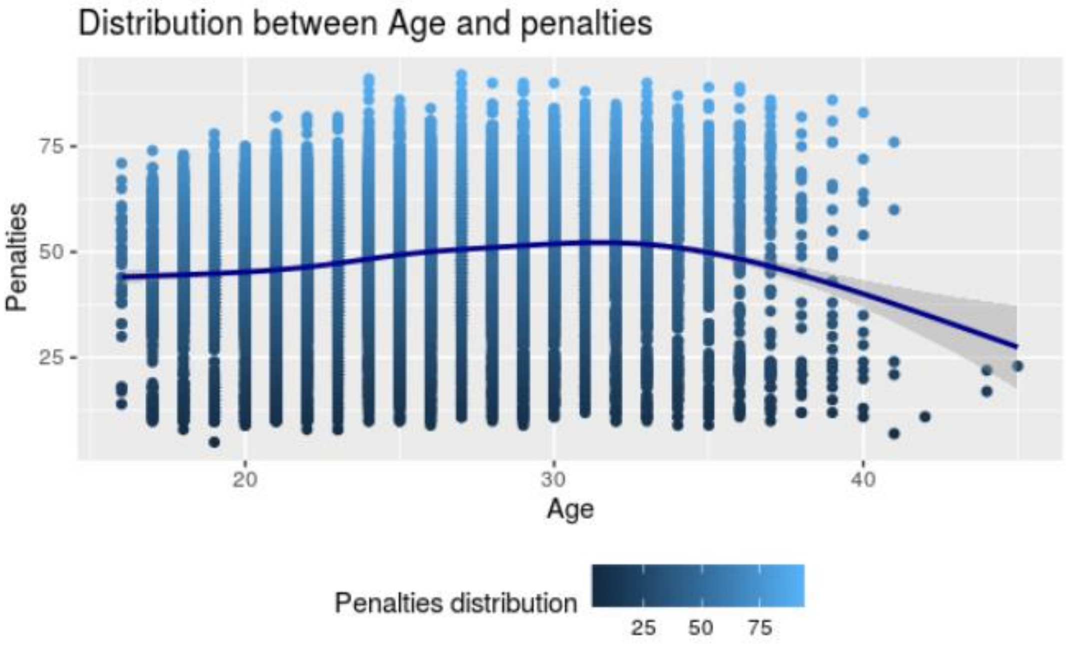


(v) Age vs Strength

On the contrary, when we have drawn the relationship between Age and strength, results are quite different than stamina which caught the eyes of the analyst. After Plotting the relationship, it is clearly visible that players' age has very less impact on its strength. It is almost stable until the mid-30s and started to decline after the late 30s. It seems quite interesting to analysts and we would like to see a relationship from a different angle which is between Age and number of Penalties to get an idea of Player's performance

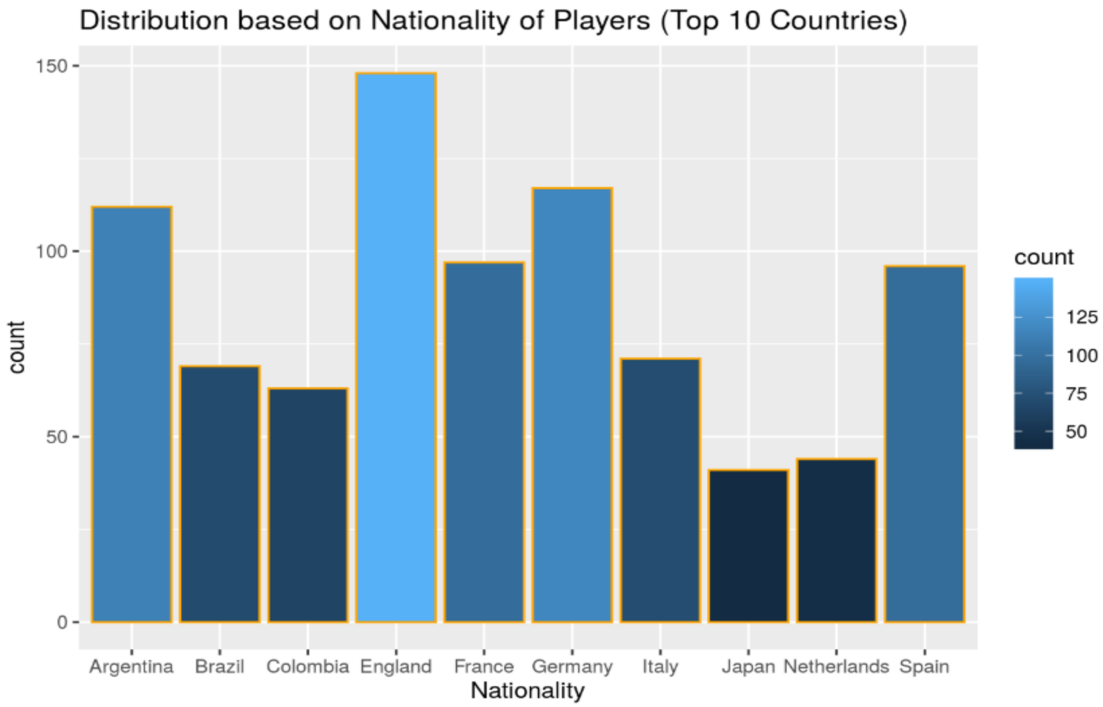
Age VS Number of Penalties:

The relationship between Penalties and Age added another dimension to the analysis. It has a more likely line with the findings when the multiple indicators of player's performance were plotted with respect to the player's age. It shows that a player's performance can not only be measured by only one indicator. The graph shows that the number of penalties are declining after the late 30s which is a good sign but it can be due to the biasedness as a number of players are less in their late 30s



(vi) Age vs Number of Penalties

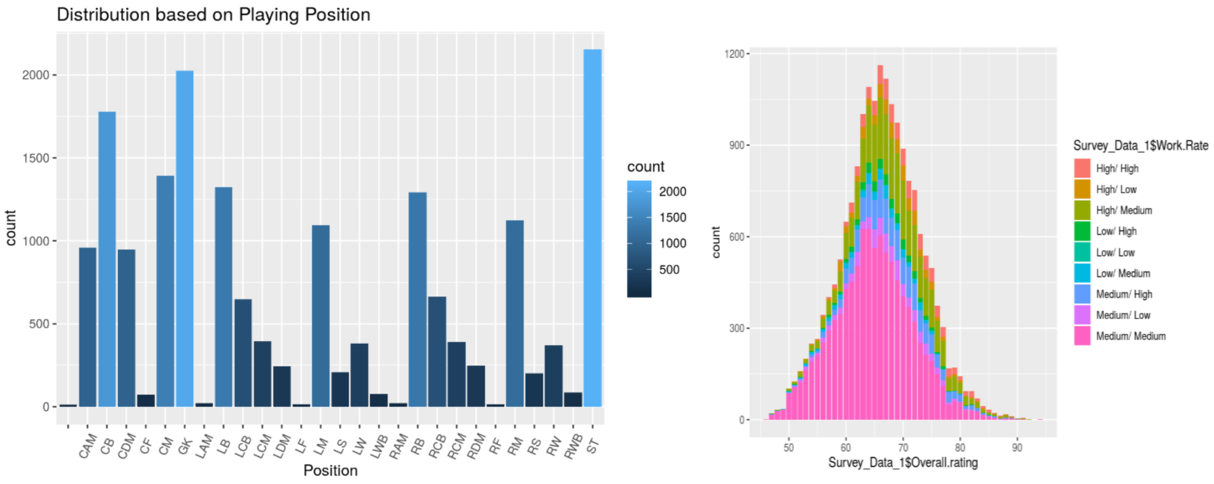
Top 10 countries – Players count wise:



(vii) Top 10 countries

We also tried to find out the distribution of players among the top 10 countries to compare the players among these nationalities. Data shows that England and Spain have most players with a proportion of 15% in overall players count. This draws interest towards this population group for further comparison of player's stats.

Distribution based on playing position and Overall rating:



(viii) Playing Position Distribution and Overall rating distribution

Lastly, we plotted the distribution among the player's playing position and which Position is preferred by most players. Data shows that Stoppers and Goal Keepers are popular among all. Since the data set is from a virtual game which justifies that people want to be more defensive to increase their probability to win a FIFA match. Moreover, Overall rating graphs shows that data is normally distributed with a mean of 66.25 and Standard deviation of 6.91

Statistical Analysis

Before statistical analysis on age, two nationality stamina comparison, and Left & right foot preference. Analysts choose the **stratified sampling strategy** as a method to filter out irrelevant data and creates equal-quantity data for comparison. We randomly selected 300 data from Spanish and 300 data points from the English player group for stamina comparison. Finally, we used the same sample for finding the goodness of fit among the work rate. Through this stratified sampling strategy, the analysis is more randomized and less selection biased. The analyst uses shuffling method as well to introduce more randomness and remove relationships to avoid biasedness

Two sample T-Test for the difference in means:

Question of Interest: Is there a difference in the average stamina between English and Spanish football players at the FIFA'19 game?
Statistical test: Two sample approach T-test.

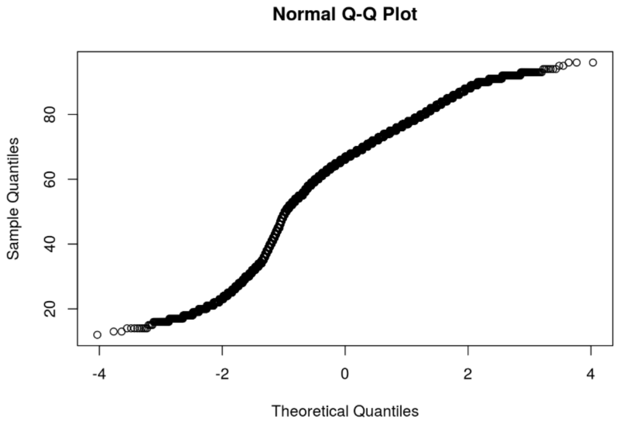
Explanation: This study focuses on the comparison of English and Spanish as two populations in terms of Stamina. The true mean stamina, as well as the true population of each group, are unknown. Therefore, the best choice of this study is a two-sample t-test.

Condition and Requirement

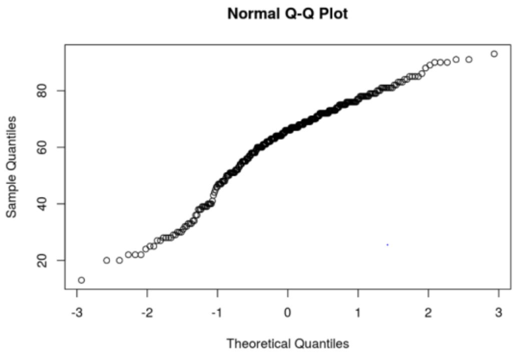
- Is the sample representative of the population?
    - Yes, Sample is representative of the population which is FIFA'19
- The question of interest has to do with the difference of means between two populations.
    - The question of interest is correlated to the difference of means stamina between English and Spanish Football player.
- 2 independent samples from 2 populations.
    - English and Spanish Football players are two independent samples from two populations for this analysis.
- The population data must be normally distributed.
    - QQ plots show both sample data are nearly normally distributed.
    - The sample size is 300 of each sample which is larger than 30.

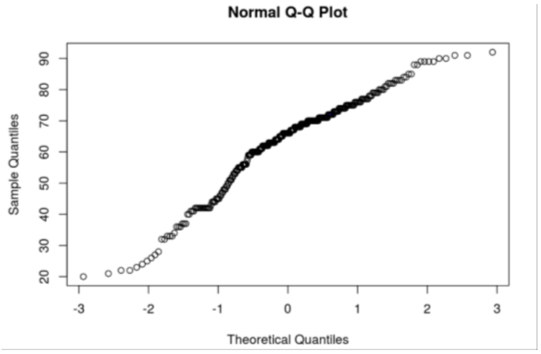
QQ - Plot:

Three Normal QQ plots indicate the data is nearly normal distributed in total sample, English football player group sample, and Spanish Football Player sample. In this study, the qqnorm() function will be utilized to plot the graph and draw a conclusion based on whether the average screen time for both populations is distributed normally or not.



(v) Total Sample





English Football player Sample

Spanish Football Player Sample

Based on the Normal Q-Q plots, the study suggests English and Spanish football player group distributions are all nearly normally distributed. Therefore, these data can be used for two-sample t-test in the next step.

Parameter:

We are interested in the true population mean difference in stamina between English and Spanish football player at FIFA'19.

Hypothesis:

$$\mu_s - \mu_e = 0$$

Null Hypothesis (H0): The true population mean Stamina for a Spanish football player is equal to the true population mean Stamina for English football player

$$\mu_s - \mu_e \neq 0$$

Alternate Hypothesis (HA): The true population mean Stamina for a Spanish football player is different from the true population mean Stamina for English football player.

Sample Statistic

Here$\overline{x_s}$ = Average Stamina of Spanish Football Player whereas $\overline{x_e}$= Average Stamina of English Football Player

$$\overline{x_s} - \overline{x_e}$$

Test Statistic

Here${S_e}^2$ = Variance of English Sample whereas ${S_s}^2$= Variance of Spanish Sample.

Also n_e= Length of English sample size and n_s= length of Spanish sample size.

$$t_{min(n_s-1,n_e-1)} = \frac{(\overline{x_s} - \overline{x_e}) - (\mu_s - \mu_e)}{\sqrt{\frac{S_e^2}{n_e} + \frac{S_s^2}{n_s}}}$$

P-Value

The analysis involves comparing the P-Value with the significance of level $\alpha = 0.05$. If P-Value is < α then our null hypothesis is rejected whereas if P-value is $> \alpha$ there exist weak evidence and we will not be able to reject the null hypothesis

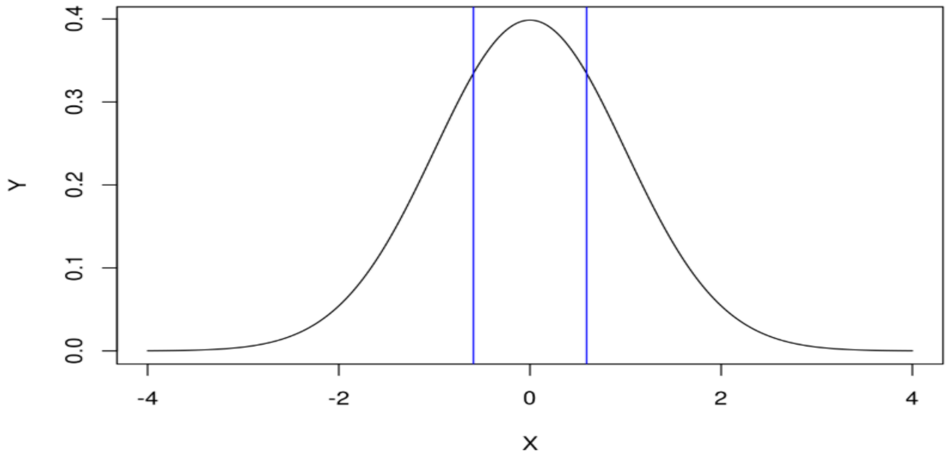
By using the t-test statistical method in R, the p-value was calculated and returned 0.5544682. The lower bound of the confidence interval is -1.751979 and the upper bound of the confidence interval is 3.258646 .

Comparison with R built-in Welch's T-test

As Per the Calculation, P-Value for both the test is almost the same

Parameters	T-test statistical method	R Built-in Welch's T-Test
P-Value	0.5544	0.5542
Lower Bound	-1.751979	-1.746932
Upper bound	3.258646	3.253599

T- test distribution graph



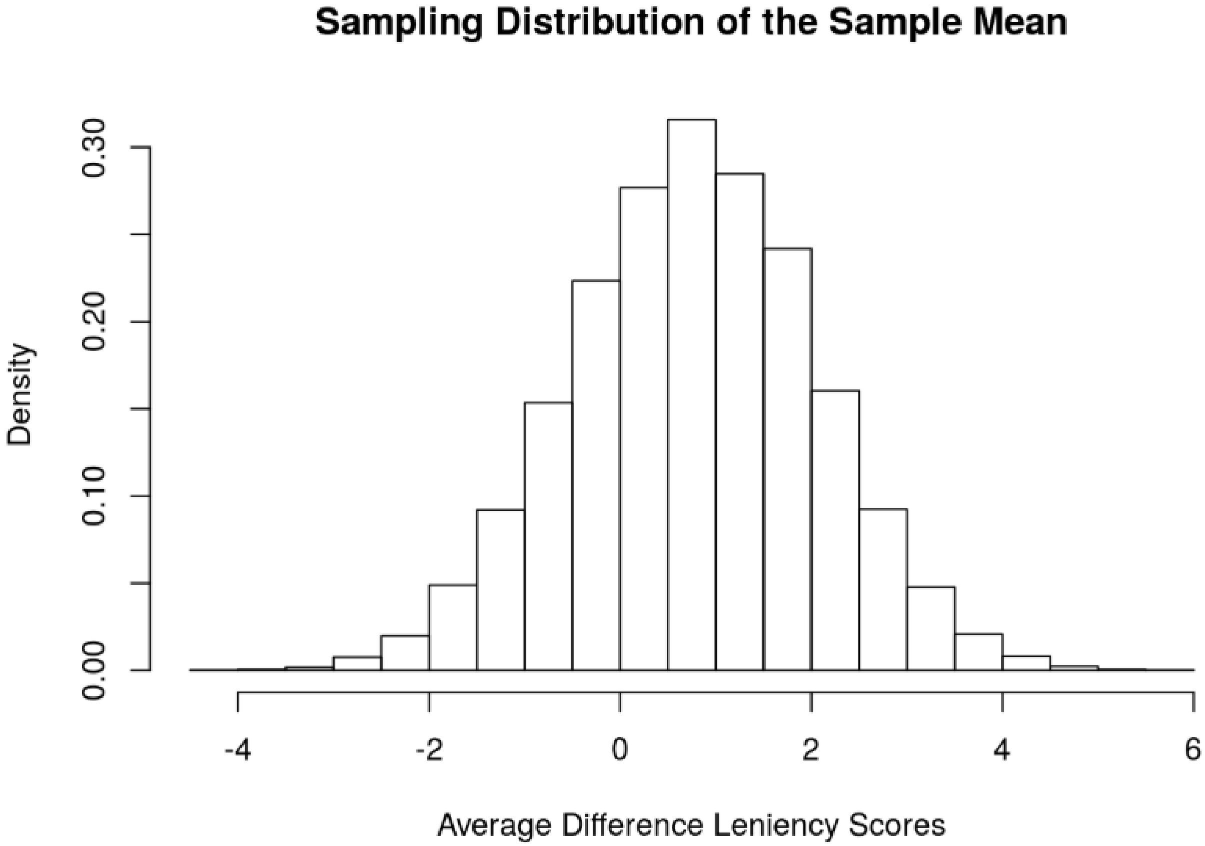
Interpretation:

From the computation, analysts found that there is no evidence (p-value=0.5542) to suggest that the true population mean Stamina for England football player is equal to the true population mean Stamina for Spain football player. We fail to reject the null hypothesis that there is no difference true mean stamina between English and Spanish football players at the level α=0.05. With 95% confidence, the true difference in mean stamina between English and Spanish football player is between -1.751979 and 3.258646

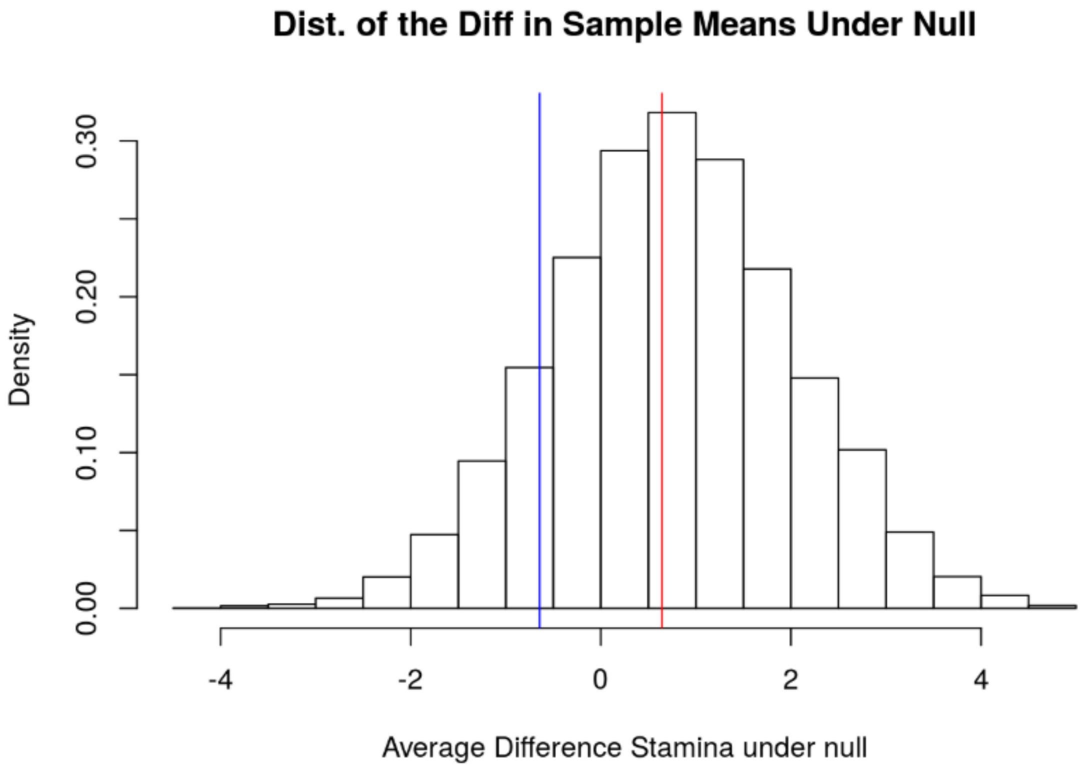
One-Sample T-Test – Bootstrapping Approach:

Sample Distribution:

We have created simulations for 10000 times for sample mean data to find sampling distribution



Sample Distribution when the Null Hypothesis is true:



Comparison with Boot Strap method

By using the Bootstrap method in R, the p-value was calculated and returned 0.6728. The lower bound of the confidence interval is -1.746932 and the upper bound of the confidence interval is 3.253599

Parameters	T-test statistical method	T – test Bootstrap method
P-Value	0.5544	0.6728
Lower Bound	-1.751979	-1.746932
Upper bound	3.258646	3.253599

One Sample T-Test – Traditional Approach:

Question of Interest:

Is Average age for Spanish player is 30 when all performance indicators are at an optimized state?

Statistical test: Traditional one sample T-test.

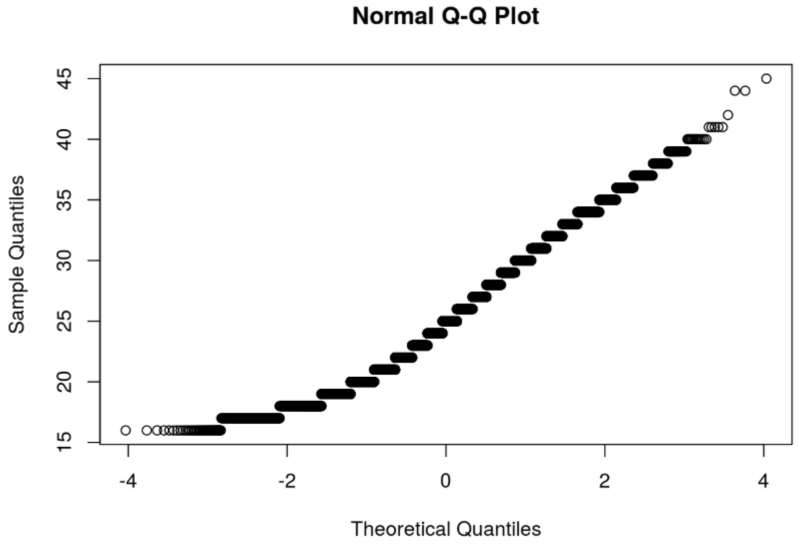
Explanation: This study focuses on finding the average age for the Spanish population. The true mean age is unknown. Therefore, the best choice of this study is a one-sample t-test.

Condition and Requirement

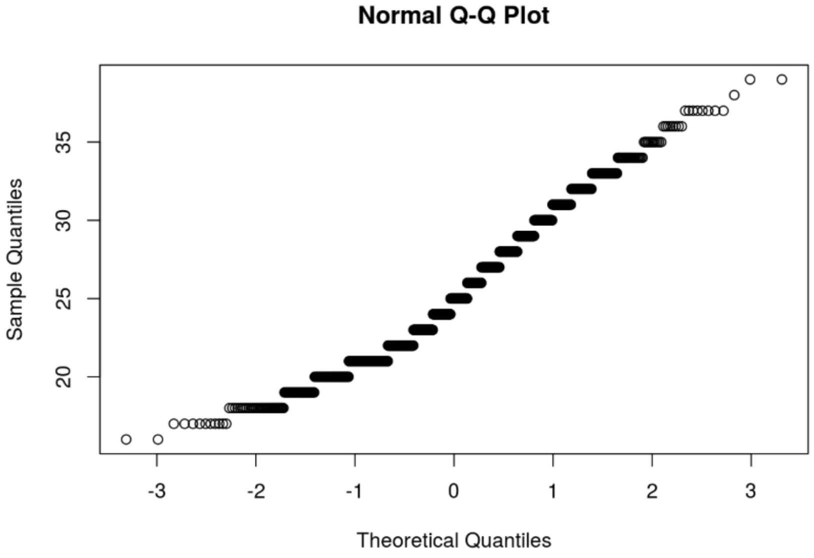
- ❖ Is the sample representative of the population
    - ➢ The sample can be representative of the population. Because study infers to find the mean age of Spanish players in FIFA’19
- ❖ One quantitative variable of interest
    - ➢ Yes, Age is Quantitative variable
- ❖ Population comes from a single population
    - ➢ Yes, data comes from a single population of FIFA'19
- ❖ The population data must be normally distributed.
    - ➢ QQ plots show both sample data are nearly normally distributed.
    - ➢ The sample size is 1071 which is larger than 30.

QQ - Plot:

Two Normal QQ plots indicate the data is nearly normal distributed in the total sample, and Spanish sample. In this study, the qqnorm() function will be utilized to plot the graph and draw a conclusion about whether data is normally distributed or not



(ix) Total Sample



(x) Spanish Sample

Parameter:

We are interested in the true population to mean age for Spanish Football player of FIFA'19

Hypothesis:

$$\mu = \mu_a = 30$$

Null Hypothesis (H0): The true population mean age for Spanish players is 30 years

$$\mu \neq \mu_a \neq 30$$

Alternate Hypothesis (HA): The true population mean age for Spanish players is different than 30 years

Sample Statistic

Here $\bar{x}$ = sample mean age

$$\bar{x}$$

Test Statistic and Distribution:

we don't know the population variance and we have to estimate it with the sample mean, the reference distribution of our test statistic shifts to a t-distribution

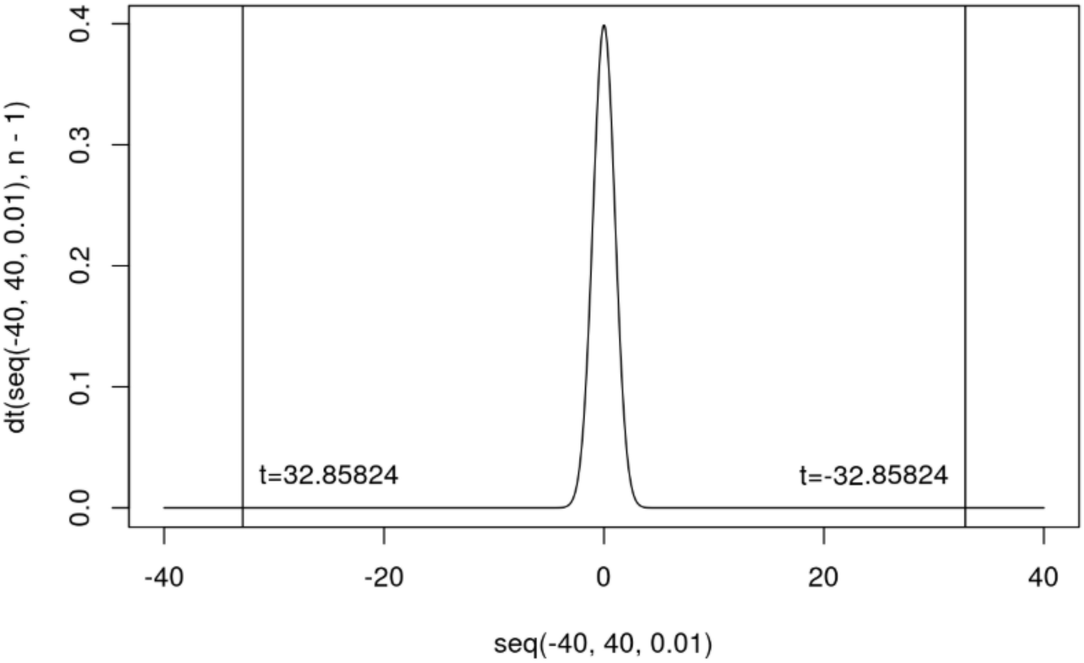
$$t_{n-1} = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$$

P- Value

The analysis involves comparing the P-Value with the significance level $\alpha = 0.05$. If P-Value is $< \alpha$ then our null hypothesis is rejected whereas if P-value is $> \alpha$ there exist weak evidence and we will not be able to reject the null hypothesis

By using the t-test statistical method in R, the p-value was calculated and returned 2.743933e-164. The lower bound of the confidence interval is 25.05169 and the upper bound of the confidence interval is 25.60938.

T-test distribution graph



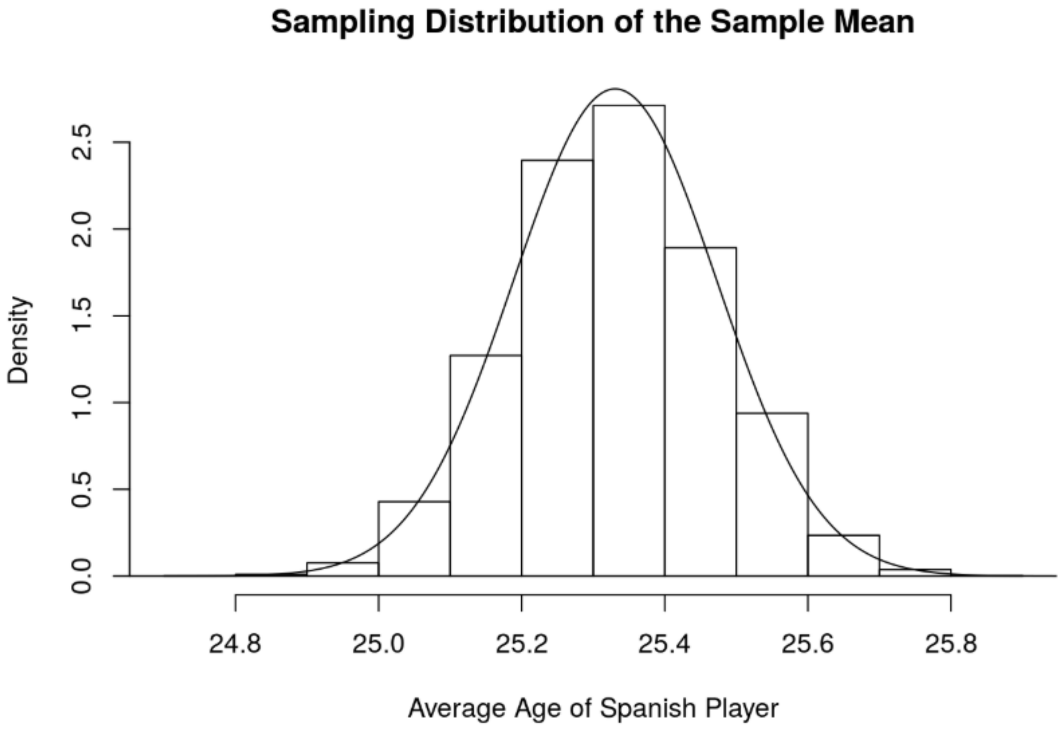
Interpretation:

There is strong evidence (p-value=2.743933e-164) to suggest that the true mean age for Spanish players is different than 30 years. We reject the null hypothesis that the true mean age for Spanish players is 30 years at the $\alpha = 0.05$ level. With 95% confidence, the true mean age is between 25.05169 years and 25.60938 years.

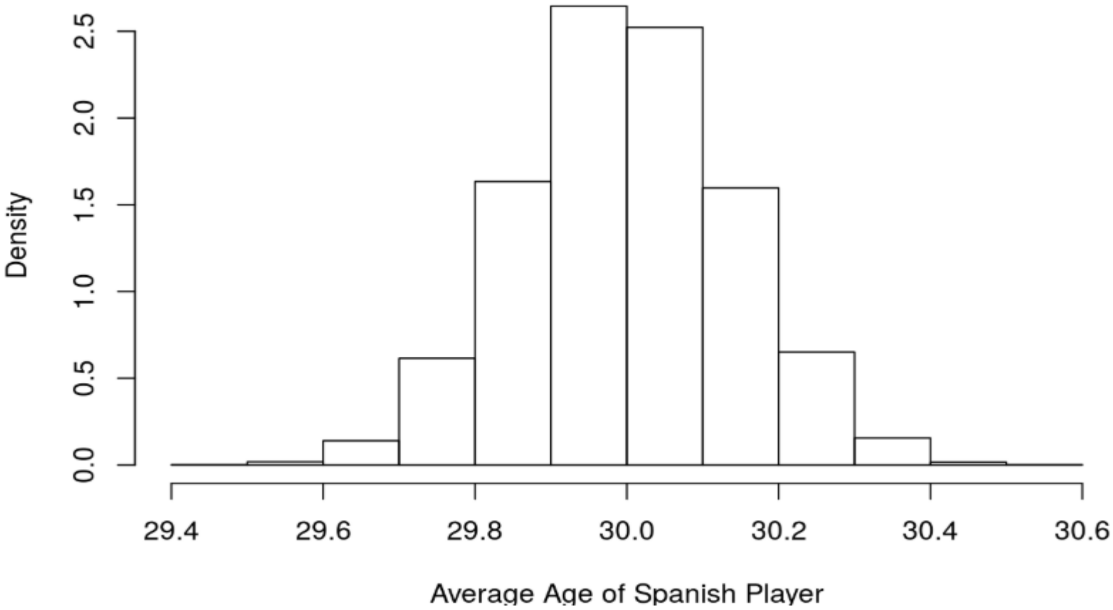
Two Sample T-Test – Bootstrapping Approach:

Sample Distribution:

We have created simulations for 10000 times for sample mean data to find sampling distribution



Sample Distribution when Null Hypothesis is true:



Comparison with Boot Strap method

By using the Bootstrap method in R, the p value was calculated and returned 0. The lower bound of the confidence interval is 25.04268 and the upper bound of the confidence interval is 25.61838.

All values are matching with traditional T method

Parameters	T-test statistical method	T – test Bootstrap method
P-Value	2.743933e-164	0
Lower Bound	25.05169	25.04268
Upper bound	25.60938	25.61838

One Sample Test of Proportion – Traditional Approach:

Question of Interest:

To find the true proportion of Left foot football player

Statistical test: Traditional one sample Test of Proportion.

Condition and Requirement

- ❖ Exact Binomial Test
  - ➢ No Requirements
- ❖ Normal Approximation
  - ➢ $n\hat{P} \geq 10$ and $n(1-\hat{P}) \geq 10$
  - ➢ (18159)(4211/18159) = 4211
  - ➢ (18159)(13948/18159) = 13948

Parameter:

The population parameter we want to make an inference to is the population proportion of football players who use the left foot.

Hypothesis:

$$\mathrm{H_0 = P_L = 0.20}$$

Null Hypothesis (H0): The true proportion of football players who use left foot is 20%

$$H_A \neq P_O \neq 0.20$$

Alternate Hypothesis (HA): The true proportion of football players who use left foot is greater than 20%

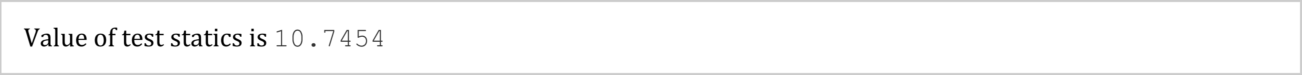
Sample Statistic

$$\hat{P} = \frac{4211}{18159} = 0.232$$

Test Statistic and Distribution:

- Exact test: there is no test statistic, find the probability directly.
- Normal approximation (using to find the test statistic - Score test statistic)

$$z = \frac{p - p_0}{\sqrt{\frac{p_0 \times (1 - p_0)}{n}}}$$



```
Value of test statics is 10.7454
```

P- Value

The analysis involves comparing the P-Value with the significance level $\alpha = 0.05$. If P-Value is $< \alpha$ then our null hypothesis is rejected whereas if P-value is $> \alpha$ there exist weak evidence and we will not be able to reject the null hypothesis

By using the t-test statistical method in R, the p-value was calculated and returned < 2.2e-16 . The lower bound of the confidence interval is 0.2267489 and the upper bound of the confidence interval is 1.0.

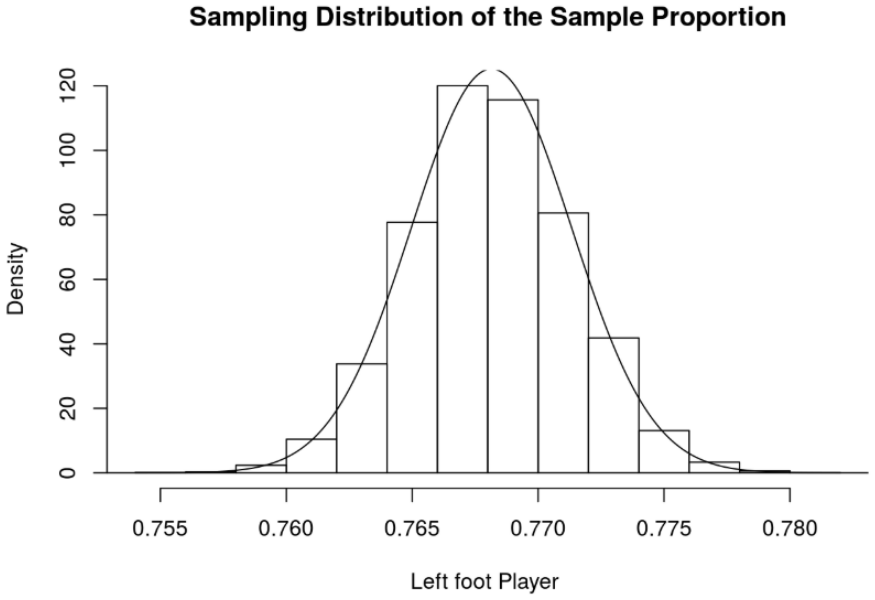
Interpretation:

Using the exact binomial methods for a one-sample test of proportion, there is Strong evidence (p-value < 0.00000000000000022) to suggest that the true proportion of left foot football player is greater than 20%. We reject the null hypothesis that the true proportion of left foot football player is equal to 20% at the $\alpha = 0.05$ level. The true proportion of left foot football player is between 0.2267489 and 1

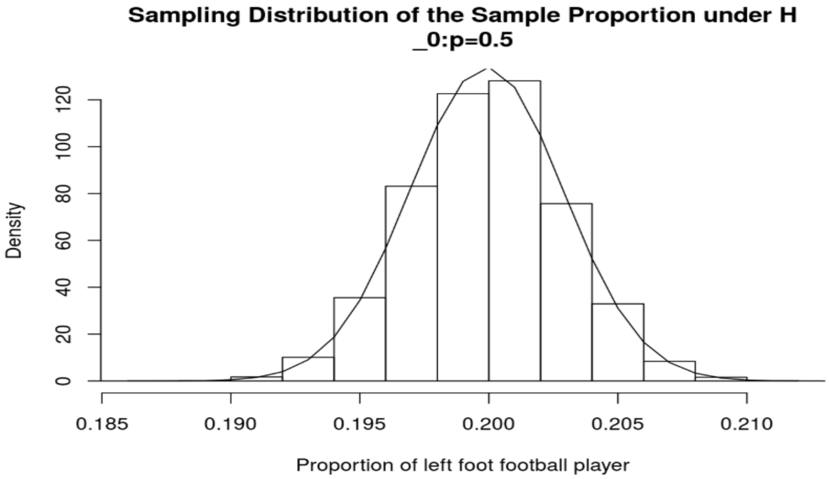
One-Sample Test of Proportion – Bootstrapping Approach:

We have created simulations for 10000 times for sample mean data to find sampling distribution

Sample Distribution:



Sample Distribution when the Null Hypothesis is true:



Comparison with Boot Strap method

By using the Bootstrap method in R, the p-value was calculated and returned 0 as the system does not calculate value below 0 whereas using the exact method result is 2.851822e-26 and normal approximation returned 3.114557e-27

All values are approximately close to each other

Parameters	T-test statistical method	T – test Bootstrap method
P-Value	< 0.00000000000000022	0
Normal Approx P-vaalue	3.114557e-27	3.114557e-27

Two sample test for difference in proportions– Traditional Approach:

Question of Interest: To find is there a difference in the true proportion of Left foot football player with a real face and right foot football player with a real face

Statistical test: Two sample test for difference in proportions

Condition and Requirement

- ❖ Sample needs to be representative of the population
    - ➢ Yes, it is
- ❖ Categorial response variable with 2 categories
    - ➢ Yes, we have 2 independent samples from 2 population
- ❖ Normal Approximation
    - ➢ $\mathrm{n}\hat{P}$≥ 10 and $\mathrm{n}\left(1-\hat{P}\right)$≥ 10

Parameter:

We are in interested in the difference between the true population proportion of real face used in left foot football players and the true population proportion of real face used in right foot football players

$$P_L - P_R$$

Hypothesis:

$$H_o: P_L - P_R = 0$$

Null Hypothesis (H0): There is no difference between the true population proportion of left foot football players with a real face and true population proportion of right foot football players with a real face

$$H_A: P_L - P_R \neq 0$$

Alternate Hypothesis (HA): There is a difference between the true population proportion of left foot football players with a real face and true population proportion of right foot football players with a real face

Sample Statistic

$$\widehat{P_L} - \widehat{P_R}$$

Test Statistic and Distribution:

Notice in this test, we're using the sample proportion for the construction of the standard error just unlike in the one-sample case where we use the null hypothesized proportion.

$$z = \frac{(\widehat{p_L} - \widehat{p_R}) - (p_L - p_R)}{\sqrt{\frac{\widehat{p_L}(1-\widehat{p_L})}{n_L} + \frac{\widehat{p_R}(1-\widehat{p_R})}{n_R}}}$$

P- Value

The analysis involves comparing the P-Value with the significance level $\alpha = 0.05$. If P-Value is $< \alpha$ then our null hypothesis is rejected whereas if P-value is $> \alpha$ there exist weak evidence and we will not be able to reject the null hypothesis

By using the t-test statistical method in R, the p-value was calculated and returned = 0.0608782. The lower bound of the confidence interval is -0.0004438872 and the upper bound of the confidence interval is 0.01988683 .

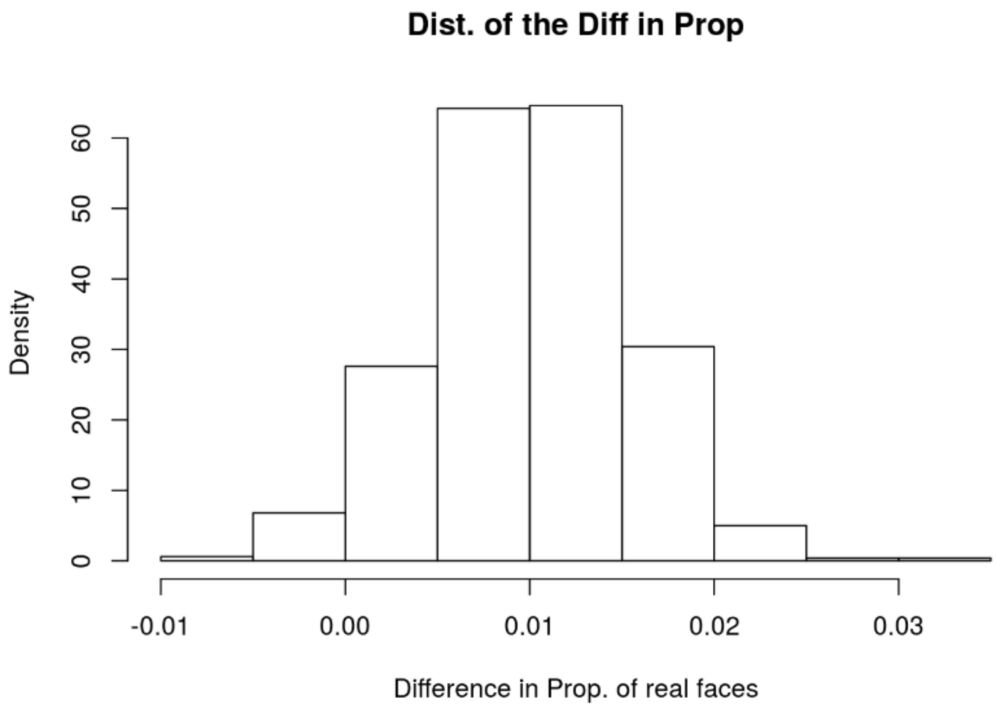
Interpretation:

Using randomization methods, there is no evidence (p-value = 0.0608782) to suggest that there is a difference between the true proportion of left foot football players with real face and true population proportion of right foot football players with real face used. We fail to reject the null hypothesis that the true population proportion of left foot football players with real face is equal to the true population proportion of right foot football players with real face at the $\alpha = 0.05$ level. The null hypothesized difference of 0 is in the confidence interval which agrees with our failure to reject the null hypothesis.

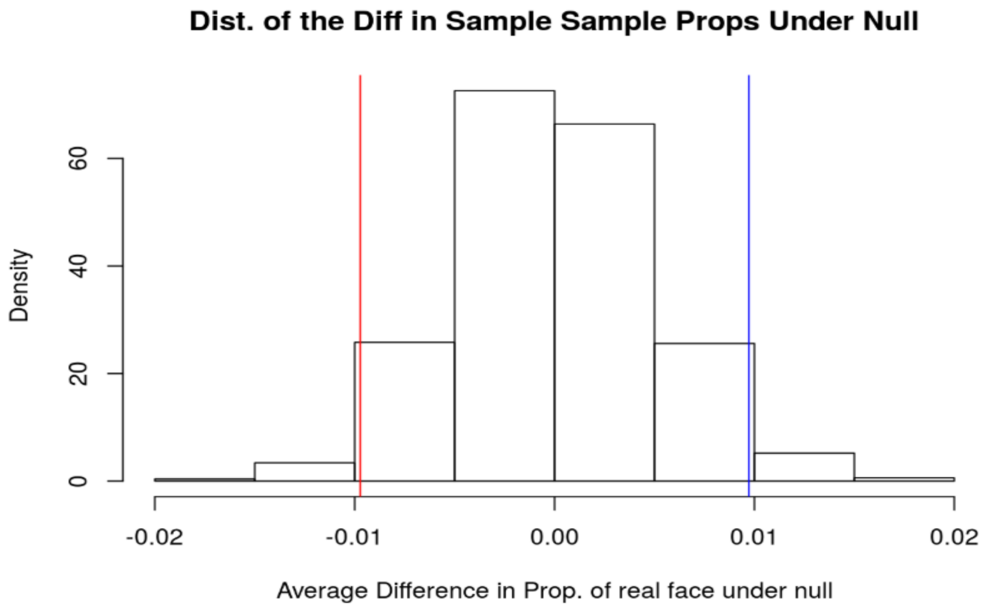
Two sample test for difference in proportions– – Bootstrapping Approach:

We have created simulations for 1000 times for sample mean data to find sampling distribution

Sample Distribution:



Sample Distribution when Null Hypothesis is true:



Comparison with Boot Strap method

By using the Bootstrap method in R, the p-value was calculated and returned = 0.057 as the system does not calculate value below 0 whereas using normal approximation returned 0.06087822

All values are approximately close to each other

Parameters	T-test statistical method	T – test Bootstrap method
P-Value	-	0.057
Normal Approx P -Value	0.0608782	0.06087822

Chi-square Goodness of Fit Test – Traditional Approach:

Question of Interest:

To find the proportion of work rate among English and Spanish football players

Statistical test: Chi-square Goodness of Fit Test

Condition and Requirement

- ❖ Single categorical variable with more than 2 categories.
    - ➢ Yes, it has 9 categories
- ❖ The expected count of each count is at least 5
    - ➢ Following categories are used:
        - ■ High/ High, High/ Low, High/ Medium, Low/ High, Low/ Low, Low/ Medium, Medium/ High, Medium/ Low, Medium/ Medium

Parameter:

We are in interested in the true

$$P_{HH}, P_{HL}, P_{HM}, P_{LH}, P_{LL}, P_{LM}, P_{MH}, P_{ML}, P_{MM}$$

Hypothesis:

$$H_o: P_{HH} = P_{HL} = P_{HM} = P_{LH} = P_{LL} = P_{LM} = P_{MH} = P_{ML} = P_{MM} = 0.02$$

Null Hypothesis (H0): The proportion of each solution choice is the same and is equal to 0.02

$$H_A: P_i \neq 0.02$$

Alternate Hypothesis (HA): The proportion of each solution choice is not equal to 0.02

There are 600 observations in our sample. If each of the solution choices had the same frequency, then each solution choice would have a count of $600 \times 0.02 = 12$

In other words, under the null hypothesis, the expected count $\mathrm{np_i} = 12$

Sample Statistic

We have 9 sample statistics

$$P_{HH}, P_{HL}, P_{HM}, P_{LH}, P_{LL}, P_{LM}, P_{MH}, P_{ML}, P_{MM}$$

Test Statistic and Distribution:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E)^2}{E} \sim \chi^2_{k-1}$$

Value of Test statistics is `9381.167`

P- Value

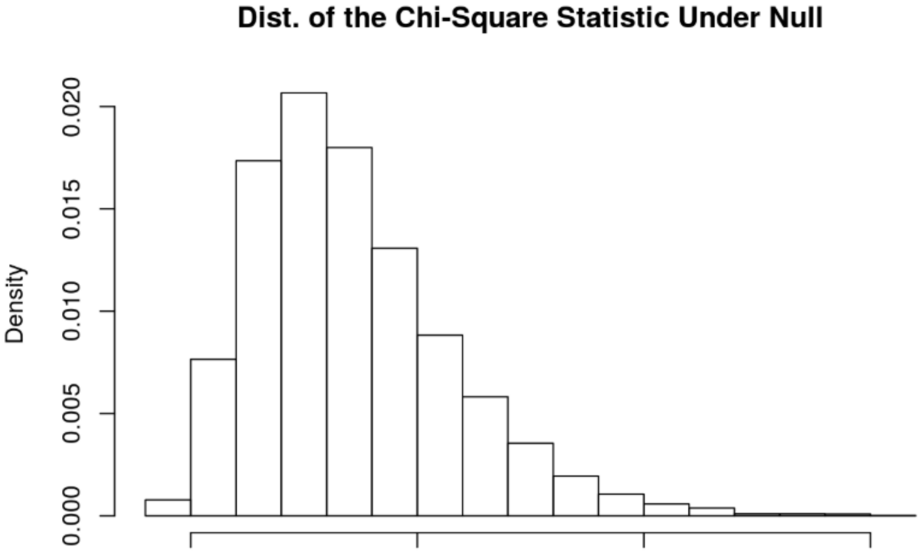
Analysis involves comparing the P-Value with the significance level $\alpha = 0.05$. If P-Value is < α then our null hypothesis is rejected whereas if P-value is $> \alpha$ there exist weak evidence and we will not be able to reject the null hypothesis

By using the t-test statistical method in R, the p-value was calculated and returned = 0.

Interpretation:

The data provides strong evidence that any of the proportions of solutions is different than 0.02. We reject the null hypothesis that the proportions of the solutions are all equal to 0.02 at the level.

Distribution of Chi-Square under Null:



Discussion

Summary and Implications:

In this observational study, analysts compared Spanish and English players as two sample groups to find if there is a difference in the true mean of stamina between Spanish and English populations in the FIFA'19 game. We have also analyzed the true proportion of left foot preferred football players with respect to the true population of 18159 data points. The study finds the distribution of sample data collected is skewed to the age group of between 20-30. The mean and median of stamina of Left foot preferred football players and Right foot preferred football player is close though right foot preferred players constitute around 78% in the sample. We have also analyzed a true proportion between left foot preferred players with real face in-game with right foot preferred players with real face in a game. Through the computation at statistical analysis, analysts find there is no evidence to suggest that the true population mean stamina for Spanish player is different than English player which cannot reject our null hypothesis but on the other hand we rejected the null hypothesis that true mean age of Spanish football player is 30 years.

This study in another way provides interesting findings related to the proportion of different work rates in Spanish and English samples. Medium/ Medium work rate constitutes about 54% of the whole sample which shows proportion is not equal for all work rate categories. We have explored the relationship between age, stamina, strength, and penalty. We found that age plays a crucial part in a player's performance

Though this study is not able to compare the overall performance between Left and right foot football players, this can be an interesting study to work on in the future. From these findings and further development in this study, analysts will be able to better understand how many football players in the game has a real face and how many players don't have a real face. Study majorly focused to find out the correlation relation between performance indicator, mean age of Spanish players and proportion of left and right foot preferred football players.

Limitation:

However, this study has its limitations. These findings on strength, Stamina and age cannot predict which team will win the match or which player we should select to make one club win the match. Overall potential rating is different along with other parameters, but this model cannot be applied to the real-world as data is purely based on a virtual game with most of its players do not even exist. Findings from the study cannot support conclusions on causation since there are many factors on which player’s performance depends.

Money plays a crucial part in a club’s performance, but this study does not concentrate on that part. Although wage, rating, club and release clause is mentioned for each player so relation can be identified if explored further.

Next Step:

There can be multiple steps which if taken can make this study more robust

1. We should consider other factors like wage, height and overall rating also in the study for establishing more identifying relationships
2. Ball control, reaction, and age are the main attributes determine a player’s potential so if we are selecting players for FIFA’19 then we should explore these areas as well

References

AN IN-DEPTH ANALYSIS OF FIFA 19. Retrieved from
https://www.dontblamethedata.com/blog/an-in-depth-analysis-of-fifa19/

Aman Shrivastava: fifa18-all-player-statistics. Retrieved from
https://github.com/amanthedorkknight/fifa18-all-player-statistics

Appendix

1. Data Set

FIFA 19 complete player dataset - https://www.kaggle.com/karangadiya/fifa19

2. Code for Test

```
# Installing Package
install.packages("ggplot2")
library(ggplot2)        # plotting & data
library(dplyr)          # data manipulation
# Loading Survey data in RStudio
Survey_Data <- read.csv("/cloud/project/fifa.csv")
summary(Survey_Data)
# Data Cleansing
Survey_Data_1 <-  Survey_Data[!(is.na(Survey_Data$Preferred.Foot) | Survey_Data$Preferred.Foot=="" ) , ]
ggplot(Survey_Data_1, aes(x=Survey_Data_1$Age, color = Survey_Data_1$Preferred.Foot)) +
  geom_histogram(fill="Grey", position="dodge", bins = 10)  +   theme(legend.position="bottom") + labs(x = "Age", color = "Legend: Preferred Foot")
ggplot(Survey_Data_1, aes(x = Survey_Data_1$Preferred.Foot , y = Survey_Data_1$Stamina)) +
    geom_boxplot() + guides(fill = TRUE) +
     theme(axis.text.x = element_text(angle = 65, vjust=0.5, hjust=0.5))+ labs(x = "Preferred Foot", y = "Stamina")
ggplot(Survey_Data_1, aes(x = Survey_Data_1$Height, color = Survey_Data_1$Preferred.Foot)) +
    geom_bar() + guides(fill = TRUE) +
     theme (legend.position="bottom") + theme (axis.text.x = element_text(angle = 65, vjust=0.5, hjust=0.5))+ labs(x = "Height" , color = "Legend: Preferred Foot")
ggplot(Survey_Data_1, aes(Position)) +
geom_bar(aes(fill = ..count..)) +
ggtitle("Distribution based on Playing Position")+
     theme(axis.text.x = element_text(angle = 65, vjust=0.5, hjust=0.5))
countries_count <- count(Survey_Data_1, Nationality)
```

```
top_10_countries <- top_n(countries_count, 10, n)

top_10_countries_1 <- top_n(countries_count,10)

top_10_country_names <- top_10_countries$Nationality


country <- filter(Survey_Data_1, Nationality == top_10_country_names)

ggplot(country, aes(x = Nationality)) +

geom_bar(col = "orange", aes(fill = ..count..)) + ggtitle("Distribution based on Nationality of Players (Top 10 Countries)") +  theme (legend.position="bottom")

g_age_overall <- ggplot(Survey_Data_1, aes(Survey_Data_1$Age, Survey_Data_1$Penalties))

g_age_overall +

geom_point(aes(color=Survey_Data_1$Penalties)) + geom_smooth(color="darkblue") +

ggtitle("Distribution between Age and penalties")  + labs(x = "Age" , y = "Penalties", color = "Penalties distribution") +  theme (legend.position="bottom")

g_age_overall <- ggplot(Survey_Data_1, aes(Survey_Data_1$Age, Survey_Data_1$Stamina))

g_age_overall +

geom_point(aes(color=Survey_Data_1$Stamina)) + geom_smooth(color="darkblue") +

ggtitle("Distribution between Age and Stamina") + labs(x = "Age" , y = "Stamina", color = "Stamina distribution") +  theme (legend.position="bottom")

g_age_overall <- ggplot(Survey_Data_1, aes(Survey_Data_1$Age, Survey_Data_1$Strength))

g_age_overall +

geom_point(aes(color=Survey_Data_1$Strength)) + geom_smooth(color="darkblue") +

ggtitle("Distribution between Age and Strength")  + labs(x = "Age" , y = "Strength", color = "Strength distribution") +  theme (legend.position="bottom")

ggplot(Survey_Data_1, aes(x = Survey_Data_1$Overall.rating, , fill = Survey_Data_1$Work.Rate)) +

  geom_bar()

set.seed(0)

Spain_1 <- filter(Survey_Data_1, Survey_Data_1$Nationality == "Spain")

Spain_2 <- Spain_1[sample(nrow(Spain_1),1071),]

summary(Spain_2)

# QQ Plot

qqnorm(Survey_Data_1$Age)
qqnorm(Spain_2$Age)
order_stats <- order(Spain_2$Age)
quantile(order_stats)
quantile(order_stats, seq(.01, .99, .1))
X_h <- mean(Spain_2$Age)
X_h
mu_0 <- 30
s <- sd(Spain_2$Age)
s
n <- length(Spain_2$Age)
n
set.seed(0)
t <- (X_h - mu_0)/(s/sqrt(n))
t
set.seed(0)
two_sided_t_pval <- pt(q = t, df = n-1, lower.tail = TRUE)*2
two_sided_t_pval
plot(seq(-40, 40, .01), dt(seq(-40, 40, .01), n-1), type="l")
# add the lines for my test statistic
abline(v=c(t, -t))
text(t,.025,"t=32.85824",srt=0.2,pos=4)
text(-t,.025,"t=-32.85824",srt=0.2,pos=2)
# lower bound
X_h+(qt(0.025, n-1)*(s/sqrt(n)))
# upper bound
X_h+(qt(0.975, n-1)*(s/sqrt(n)))
set.seed(0)
# This data is pretty skewed so even though n is large, I'm going to do a lot of simulations
num_sims <- 1000
# A vector to store my results
results <- rep(NA, num_sims)
# A loop for completing the simulation
for(i in 1:num_sims){
 results[i] <- mean(sample(x = Spain_2$Age,
 size = n,
 replace = TRUE))
}
# Finally plot the results
hist(results, freq = FALSE, main='Sampling Distribution of the Sample Mean', x
lab = 'Average Age of Spanish Player', ylab = 'Density')
# estimate a normal curve over it - this looks pretty good!
lines(x = seq(24, 26, .001), dnorm(seq(24, 26, .001), mean = X_h, sd = s/sqrt(
n)))
set.seed(0)
# Shift the sample so that the null hypothesis is true
time_given_H0_true <- Spain_2$Age - mean(Spain_2$Age) + mu_0
# This data is pretty skewed so even though n is large, I'm going to do a lot
of simulations
num_sims <- 1000
# A vector to store my results
results_given_H0_true <- rep(NA, num_sims)
# A loop for completing the simulation
for(i in 1:num_sims){
 results_given_H0_true[i] <- mean(sample(x = time_given_H0_true,
 size = n,
 replace = TRUE))
}
results_given_H0_true
# Finally plot the results
hist(results_given_H0_true, freq = FALSE, main='Sampling Distribution of the S
ample Mean, Given Null Hypothesis is True', xlab = 'Average Age of Spanish Pla
yer', ylab = 'Density')
# add line to show values more extreme on upper end
abline(v=X_h, col = "red")
# add line to show values more extreme on lower end
low_end_extreme <- mean(results_given_H0_true)+(mean(results_given_H0_true)-X_
h)
abline(v=low_end_extreme, col="red")
set.seed(0)
```

```
# counts of values more extreme than the test statistic in our original sample
, given H0 is true
# two sided given the alternate hypothesis
count_of_more_extreme_lower_tail <- sum(results_given_H0_true <= low_end_extre
me)
count_of_more_extreme_upper_tail <- sum(results_given_H0_true >= X_h)
bootstrap_pvalue <- (count_of_more_extreme_lower_tail - count_of_more_extreme_
upper_tail)/1000
bootstrap_pvalue
# two sided t p-value
two_sided_t_pval
# need the standard error which is the standard deviation of the results
bootstrap_SE_X_bar <- sd(results)
# an estimate is to use the formula statistic +/- 2*SE
c(X_h - 2*bootstrap_SE_X_bar, X_h + 2*bootstrap_SE_X_bar)
# compare to our t-methods
c(X_h+(qt(0.025, n-1)*(s/sqrt(n))), X_h+(qt(0.975, n-1)*(s/sqrt(n))))
set.seed(0)
summary(Survey_Data_1$Preferred.Foot)
Survey_Data_2 <- Survey_Data_1[sample(nrow(Survey_Data_1),100),]
summary(Survey_Data_2$Preferred.Foot)
n <- 18159
x <- 4211
p_hat <- 4211/18159
p_hat
z <- (p_hat - .20) / sqrt((.20*(1-.20)) / n)
z
binom.test(x=4211, n = 18159, p=(.20), alternative="greater")
pnorm(z, lower.tail = FALSE)
cat("exact binomial test")
binom.test(x=4211, n = 18159, p=(.20), alternative="greater")$conf.int
cat("normal approx")
c(.23 - (1.64)*sqrt(((.23)*(1-.23))/n), 1)
Left_Football <- Survey_Data_1$Preferred.Foot
Left_Football
```

```
levels(Left_Football) <- c(1,0)

Left_Football

table(Left_Football)

# This data is pretty skewed so even though n is large, I'm going to do a lot
of simulations

set.seed(0)

num_sims <- 10000

# A vector to store my results

results <- rep(NA, num_sims)

# A loop for completing the simulation

for(i in 1:num_sims){

 results[i] <- mean(as.numeric(sample(x = Left_Football ,

 size = n,

 replace = TRUE))-1)

}

# Finally plot the results

hist(results, freq = FALSE, main='Sampling Distribution of the Sample Proporti
on', xlab = 'Left foot Player ', ylab = 'Density')

# estimate a normal curve over it - this looks pretty good!

lines(x = seq(0.10, 1.0000000, .0001), dnorm(seq(0.10, 1.0000000, .0001), mean
= mean(results), sd = sd

(results)))


cat("Bootstrap Confidence Interval")

c(quantile(results, c(.05, 1)))

cat("exact binomial test")

binom.test(x=4211, n = 18159, p=(.20), alternative="greater")$conf.int

cat("normal approx")

c(p_hat - (1.64)*sqrt(((p_hat)*(1-p_hat))/n), 1)

set.seed(0)

# Under the assumption that the null hypothesis is true,

Left_Football_new <- rep(c(1, 0), c(0.20*n, (1-0.20)*n))

num_sims <- 10000

# A vector to store my results

results <- rep(NA, num_sims)
```

```
# A loop for completing the simulation
for(i in 1:num_sims){
 results[i] <- mean(sample(x = Left_Football_new,
 size = n,
 replace = TRUE))
}
# Finally plot the results
hist(results, freq = FALSE, main='Sampling Distribution of the Sample Proporti
on under H
_0:p=0.5', xlab = 'Proportion of left foot football player', ylab = 'Density')
# estimate a normal curve over it - this looks pretty good!
lines(x = seq(.10, .30, .001), dnorm(seq(.10, .30, .001), mean = mean(results)
, sd = sd
(results)))
abline(v=p_hat, col="red")
count_of_more_extreme_upper_tail <- sum(results >= p_hat)
bootstrap_pvalue <- count_of_more_extreme_upper_tail/num_sims
cat("Bootstrap p-value")
bootstrap_pvalue
cat("Exact Binomial p-value")
binom.test(x = 4211, n = 18159, p = 0.20, alternative = "greater")$p.value
cat("Normal Approximation p-value")
pnorm(z, lower.tail = FALSE)
set.seed(0)
England_3 <- filter(Survey_Data_1, Survey_Data_1$Nationality == "England")
England_4 <- England_3[sample(nrow(England_3),300),]

Spain <- filter(Survey_Data_1, Survey_Data_1$Nationality == "Spain")
Spain <- Spain[sample(nrow(Spain),300),]

sample_data <- rbind(Spain,England_4)
# QQ Plot
qqnorm(Survey_Data$Stamina)
qqnorm(England_4$Stamina)
qqnorm(Spain$Stamina)
```

```
# Sample Mean
X_bar_e <- mean(England_4$Stamina)
X_bar_s <- mean(Spain$Stamina)
X_bar_e
X_bar_s
# Sample Variance
s_e <- sd(England_4$Stamina)**2
s_s <- sd(Spain$Stamina)**2
s_e
s_s
# Sample Size
n_e <- length(England_4$Stamina)
n_s <- length(Spain$Stamina)
n_e
n_s
#Null Hypothises
mu <- 0
# T Test
t <- (X_bar_s - X_bar_e - mu)/sqrt((s_s/n_s) + (s_e/n_e))
t
# p-value for two sided upper
two_sided_diff <- pt(q=t, df = min(n_s-1, n_e-1), lower.tail = FALSE) * 2
two_sided_diff
Alpha <- 0.05
Confidence_Interval <- 0.95
# Lower Bound
L_bound <- (X_bar_s - X_bar_e) + (qt(0.025, min(n_s, n_e)-1)* sqrt((s_s/n_s) +
(s_e/n_e)))
L_bound
# Upper Bound
U_bound <- (X_bar_s - X_bar_e) + (qt(0.975, min(n_s, n_e)-1)* sqrt((s_s/n_s) +
(s_e/n_e)))
U_bound
# R built in t-test function
t.test(Spain$Stamina, England_4$Stamina)
```

```
# Histogram of the sampling distribution
mu <- mean(sample_data$Stamina)
sd <- sd(sample_data$Stamina)
h <- hist(sample_data$Stamina, xlim = c(0,100), xlab = 'Average Stamina', main
= 'Histogram of sampling distribution')
lb <- mu - 1.96*sd
ub <- mu + 1.96*sd
abline(v = c(mu, lb, ub), lty = 2)
x_axis <- seq(min(sample_data$Stamina),max(sample_data$Stamina),length=600)
y_axis <- dnorm(x_axis, mu, sd)*length(x_axis)
lines(x_axis, y_axis, col = "blue")
# T- Test distribution graph
n <- min(n_e, n_s)
X <- seq(-4, 4, .01)
Y <- dt(X, n-1)
plot(X, Y, type = 'l')
abline(v = c(t, -t),  col = "blue")
# Confidence Interval graph
plot(X, Y, type = 'l')
abline(v = qnorm(0.975), col = "Green")
abline(v = qnorm(0.025), col = "Green")
abline(v = 0, col = "black")
# Difference between normal distribution & T distribution
plot(X,Y,type = 'l')
lines(X,dnorm(X), col = 'yellow')
set.seed(0)
num_sims <- 10000
# A vector to store my results
results <- rep(NA, num_sims)
# A loop for completing the simulation
for(i in 1:num_sims){
mean_Spain <- mean(sample(x = Spain$Stamina,size = 300,
 replace = TRUE ))
 mean_England_4 <- mean(sample(x = England_4$Stamina, size = 300,
 replace = TRUE))
 results[i] <- mean_Spain - mean_England_4
}
# Finally plot the results
hist(results, freq = FALSE, main='Sampling Distribution of the Sample Mean', x
lab = 'Average Difference Leniency Scores', ylab = 'Density')


# Bootstrap one-sided CI
c(quantile(results, c(.975, .025)))


#compare to our t-methods
t.test(Spain$Stamina,England_4$Stamina)$conf.int
set.seed(0)
num_sims <- 10000
# A vector to store my results
results_given_H0_true <- rep(NA, num_sims)
# A loop for completing the simulation
for(i in 1:num_sims){
 # idea here is if there is no relationshipm we should be able to shuffle the
groups
#shuffled_groups <- transform(sample_data, Group=sample_data(Group))
 mean_England_4 <- mean(sample(x = England_4$Stamina,
 replace = TRUE))
 mean_Spain <- mean(sample(x = Spain$Stamina,
 replace = TRUE))
 results_given_H0_true[i] <- mean_Spain - mean_England_4
}
results_given_H0_true
# Finally plot the results
hist(results_given_H0_true, freq = FALSE,
 main='Dist. of the Diff in Sample Means Under Null',
 xlab = 'Average Difference Stamina under null',
 ylab = 'Density')
diff_in_sample_means <- mean_Spain - mean_England_4
abline(v=diff_in_sample_means, col = "blue")
```

```r
abline(v=abs(diff_in_sample_means), col = "red")
diff_in_sample_means
set.seed(0)
# counts of values more extreme than the test statistic in our original sample, given H0is true
# two sided given the alternate hypothesis
count_of_more_extreme_lower_tail <- sum(results_given_H0_true <= diff_in_sample_means)
count_of_more_extreme_upper_tail <- sum(results_given_H0_true >= abs(diff_in_sample_means))
bootstrap_pvalue <- (count_of_more_extreme_lower_tail + count_of_more_extreme_upper_tail)/num_sims
count_of_more_extreme_lower_tail
count_of_more_extreme_upper_tail
cat("Bootstrap p-value")
bootstrap_pvalue
#t-test p-value
t.test(Spain$Stamina,England_4$Stamina)$p.value

Left_foot <- filter(Survey_Data_1, Survey_Data_1$Preferred.Foot == "Left")
summary(Left_foot$Real.Face)
Left_foot <- Left_foot[sample(nrow(Left_foot),4211),]

Right_foot <- filter(Survey_Data_1, Survey_Data_1$Preferred.Foot == "Right")
summary(Right_foot$Real.Face)
Right_foot <- Right_foot[sample(nrow(Right_foot),13948),]

sample_data_1 <- rbind(Left_foot,Right_foot )
# the parts of the test statistic
# sample props
p_hat_L <- 415/4211
p_hat_L
p_hat_R <- 1239/13948
p_hat_R
# null hypothesized population prop difference between the two groups
```

```
p_0 <- 0
# sample size
n_l <- 4211
n_r <- 13948
# sample variances
den_p_l <- (p_hat_L*(1-p_hat_L))/n_l
den_p_r <- (p_hat_R*(1-p_hat_R))/n_r
# z-test test statistic
z <- (p_hat_L - p_hat_R - p_0)/sqrt(den_p_l + den_p_r)
z
# two sided p-value
two_sided_diff_prop_pval <- pnorm(q = z, lower.tail = FALSE)*2
two_sided_diff_prop_pval
# lower bound
(p_hat_L - p_hat_R)+(qnorm(0.025)*sqrt(den_p_l + den_p_r))

# upper bound
(p_hat_L - p_hat_R)+(qnorm(0.975)*sqrt(den_p_l + den_p_r))
set.seed(0)
# Make the data
Left <- rep(c(1, 0), c(415, n_l - 415))
Right <- rep(c(1,0), c(1239, n_r -1239))
num_sims <- 1000
# A vector to store my results
results <- rep(NA, num_sims)
# A loop for completing the simulation
for(i in 1:num_sims){
 prop_war <- mean(sample(Left,
 size = n_l,
 replace = TRUE))
 prop_opps <- mean(sample(x = Right,
 size = n_r,
 replace = TRUE))
 results[i] <- prop_war - prop_opps
}

# Finally plot the results

hist(results, freq = FALSE, main='Dist. of the Diff in Prop', xlab = 'Differen
ce in Prop. of real faces', ylab = 'Density')

cat("Bootstrap")

c(quantile(results, c(.025, .975)))

cat("Normal Approximation")

c((p_hat_L - p_hat_R)+(qnorm(0.025)*sqrt(den_p_l + den_p_r)), (p_hat_L - p_hat
_R)+(qnorm(0.975)*sqrt(den_p_l + den_p_r)))


# Make the data

df_combined <- data.frame("Real Face used" = c(Left, Right),

 "team" = rep(c("Left Foot footballers", "Right Foot footballers"), c(n_l, n_r
)))

# Sanity checks

summary(df_combined$team)

mean(df_combined$Real.Face.used[df_combined$team=="Left Foot footballers"]) ==
p_hat_L

mean(df_combined$Real.Face.used[df_combined$team=="Right Foot footballers"]) =
= p_hat_R

set.seed(0)

num_sims <- 1000

# A vector to store my results

results_given_H0_true <- rep(NA, num_sims)

# A loop for completing the simulation

for(i in 1:num_sims){

 # idea here is if there is no relationshipm we should be able to shuffle the
groups

 shuffled_groups <- transform(df_combined, team=sample(team))

 prop_warriors <- mean(shuffled_groups$Real.Face.used[shuffled_groups$team=="L
eft Foot footballers"

])

 prop_opponents <- mean(shuffled_groups$Real.Face.used[shuffled_groups$team=="
Right Foot footballers"])

 results_given_H0_true[i] <- prop_warriors - prop_opponents

}

# Finally plot the results
```

```
hist(results_given_H0_true, freq = FALSE,
 main='Dist. of the Diff in Sample Sample Props Under Null',
 xlab = 'Average Difference in Prop. of real face under null',
 ylab = 'Density')

diff_in_sample_props <- p_hat_L - p_hat_R

abline(v=diff_in_sample_props, col = "blue")

abline(v=-diff_in_sample_props, col = "red")

set.seed(0)

# counts of values more extreme than the test statistic in our original sample
, given H0 is true

# two sided given the alternate hypothesis

count_of_more_extreme_lower_tail <- sum(results_given_H0_true <= -diff_in_samp
le_props)

count_of_more_extreme_upper_tail <- sum(results_given_H0_true >= diff_in_sampl
e_props)

bootstrap_pvalue <- (count_of_more_extreme_lower_tail + count_of_more_extreme_
upper_tail)/num_sims

cat("Bootstrap p-value")

bootstrap_pvalue

cat("Normal Approx p-value")

two_sided_diff_prop_pval

summary(sample_data$Work.Rate)

head(sample_data$Work.Rate)

table(sample_data$Work.Rate)

prop.table(table(sample_data$Work.Rate))

sum(((table(sample_data$Work.Rate) - 12)^2)/12)

pchisq(9381.167, df = 9-1, lower.tail = FALSE)

# Create our data under the assumption that H_0 is true

solutions_under_H_0 <- rep(c("HH", "HL", "HM", "LH", "LL", "LM", "MH", "ML", "
MM"), 12)

# Sanity Check

table(solutions_under_H_0)

set.seed(0)

num_sims <- 10000

# A vector to store my results

chisq_stats_under_H0 <- rep(NA, num_sims)
```

```r
# A loop for completing the simulation
for(i in 1:num_sims){
 new_samp <- sample(solutions_under_H_0, 600, replace = T)
 chisq_stats_under_H0[i] <- sum(((table(new_samp) - 12)^2)/12)
}

# What do you notice about the distribution of this statistic?

hist(chisq_stats_under_H0, freq = FALSE,
 main='Dist. of the Chi-Square Statistic Under Null',
 xlab = 'Chi-Square Stat under Null',
 ylab = 'Density')
abline(v=sum(((table(sample_data$Work.Rate) - 12)^2)/12), col="red")
sum(chisq_stats_under_H0 >= sum(((table(sample_data$Work.Rate)-12)^2)/12))/num_sims
```