Milestone 2 Final Project

Mohammed Saif Wasay (002815958)

MPS in Informatics, Northeastern University

ALY 6010: Probability Theory and Introductory Statistics

Harpreet Sharma

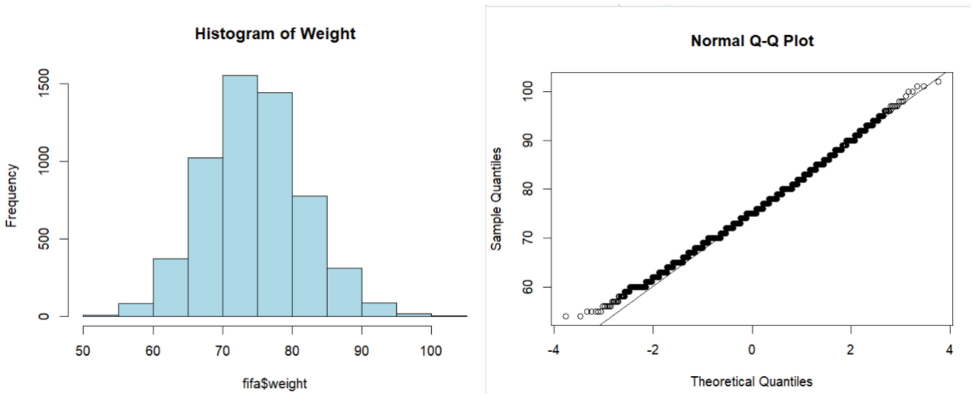
June 15th, 2024

Introduction

In this assignment milestone, we will examine the FIFA dataset on which we had performed data cleaning and exploratory data analysis in the previous milestone, and perform hypothesis testing and with the help of inferential statistics answer a few questions related to the data. The data has a collection of approximately 5600 rows and two major category of players that is Goalkeepers and Outfielders. There are columns for players height, weight, age and many more. On these parameters, we will run a few hypothesis tests.

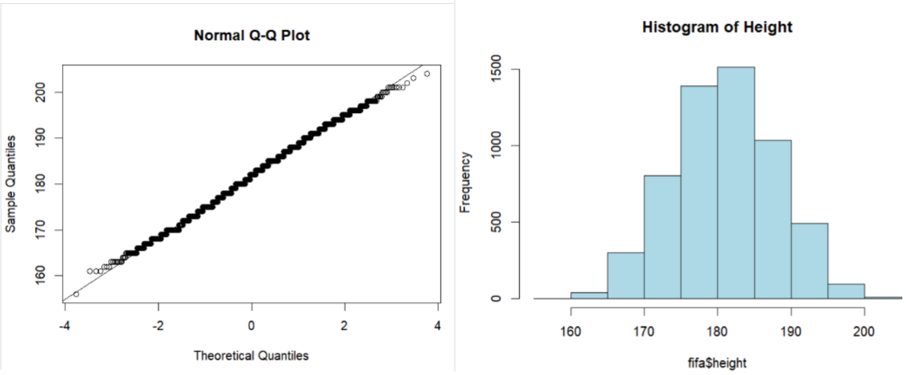
Initially we clean the data as done in the previous milestone. Before we begin our hypothesis testing, we check for the normality of our data. There are multiple ways by which we can check normality of data, for this project we plot a histogram and a Q-Q plot to check if the height and weight of players are normally distributed.

Figure 1: Normality of Player Weight



With reference to the histogram and Q-Q plot, we can see that the player weight distribution is approaching to be normal, there are few outliers that effect the data normality.

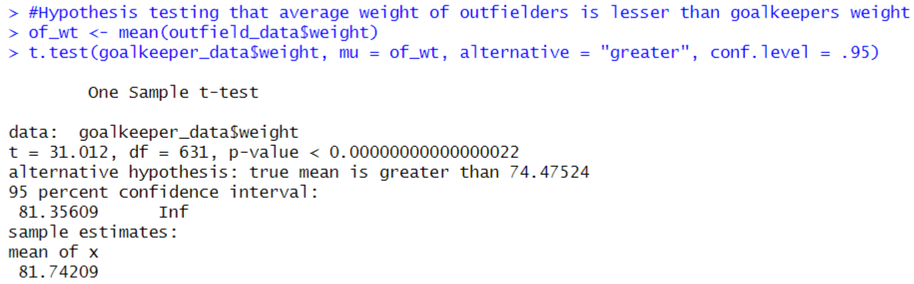
Figure 2: Normality of Player Height



With reference to the histogram and Q-Q plot, we can see that the player height distribution is approaching to be normal, there are few outliers that effect the data normality. In Q-Q plot we can see data points following closely to theoretical line.

On the data, we ran three hypothesis tests. The questions that we are looking to answer are: Is there a statistical difference in the average weight of the outfielders and goalkeepers? Are goalkeepers statistically taller than outfielders? Usually in soccer goalkeepers tend to have longer careers than outfielders, is this statistically true that average age of goalkeepers are more than outfielders?

Figure 3: Testing the hypothesis on players' weight

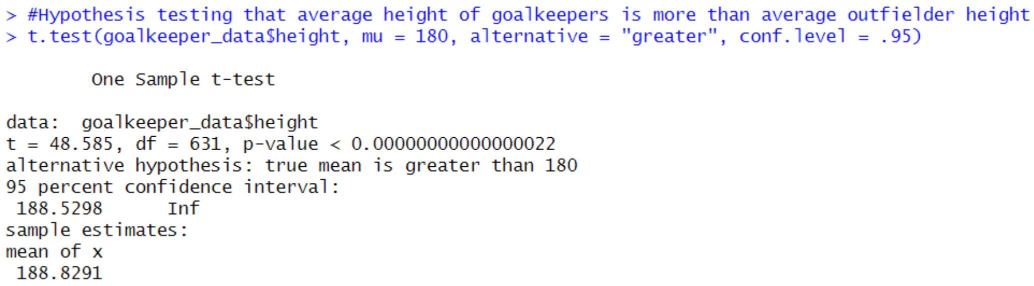


```
> #Hypothesis testing that average weight of outfielders is lesser than goalkeepers weight
> of_wt <- mean(outfield_data$weight)
> t.test(goalkeeper_data$weight, mu = of_wt, alternative = "greater", conf.level = .95)

        One Sample t-test

data:  goalkeeper_data$weight
t = 31.012, df = 631, p-value < 0.00000000000000022
alternative hypothesis: true mean is greater than 74.47524
95 percent confidence interval:
 81.35609      Inf
sample estimates:
mean of x 
 81.74209 
```

For our first hypothesis test, we are checking that the average weight of the goalkeepers is greater than that of outfielders, and we choose One Sample t-test with 95% confidence interval as it provides a balance between precision and certainty. The null hypothesis here is that there is no significant difference between average weight of goalkeepers and outfielders. Alternate hypothesis is that the average goalkeeper weight is greater than outfielder weight that is 74.47 kg.

The t-value comes around 31.012, degree of freedom is 631 and p-value of 0.00000000000000022. The p-value is extremely small, well below conventional significance levels (e.g., 0.05). This suggests strong evidence against the null hypothesis. In practical terms, it indicates that the likelihood of observing such an extreme result if the true mean weight were 74.47524 kg is exceedingly low. The 95% confidence interval for the true mean includes infinity and 81.35609 kg. In summary, based on this one-sample t-test, there is sufficient evidence to reject the null hypothesis and the statistical analysis strongly supports the claim that the true mean weight of goalkeepers is significantly greater than 74.47524 kg.

Figure 4: Testing the hypothesis on players' height

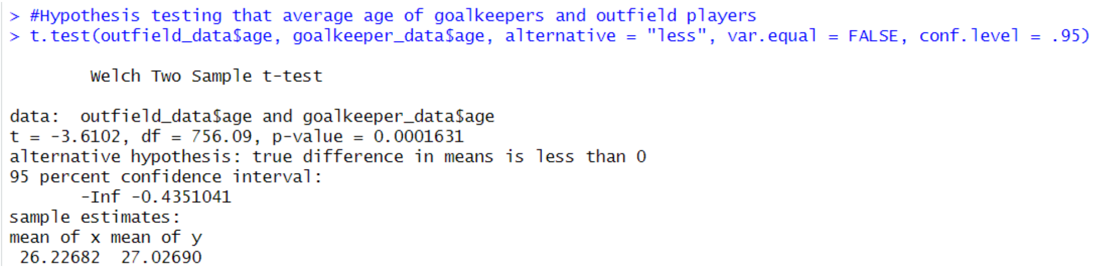


```
> #Hypothesis testing that average height of goalkeepers is more than average outfielder height
> t.test(goalkeeper_data$height, mu = 180, alternative = "greater", conf.level = .95)

        One Sample t-test

data:  goalkeeper_data$height
t = 48.585, df = 631, p-value < 0.00000000000000022
alternative hypothesis: true mean is greater than 180
95 percent confidence interval:
 188.5298      Inf
sample estimates:
mean of x 
 188.8291 
```

For height we are testing that the average height of the goalkeepers is greater than 180 cm and we choose 95% confidence interval as it provides a balance between precision and certainty. The null hypothesis here is that there is no statistical difference between goalkeeper and outfielder's heights and alternative hypothesis is that the average goalkeepers heights are greater than average outfielder height.

The p-value is extremely small, well below conventional significance levels (e.g., 0.05). This suggests strong evidence against the null hypothesis. In practical terms, it indicates that the likelihood of observing such an extreme result if the true mean height were 180 is exceedingly low. The confidence interval for the mean height is calculated to be between 188.5298 and positive infinity. This range indicates that, with 95% confidence, the true mean height is expected to be greater than 188.5298. The statistical analysis strongly supports the claim that the true mean height of goalkeepers is significantly greater than 180. The evidence is robust, considering both the large t-statistic and the extremely small p-value. The 95% confidence interval further reinforces the conclusion, suggesting a substantial difference in height for goalkeepers.

Figure 5: Testing of the player's age hypothesis



```
> #Hypothesis testing that average age of goalkeepers and outfield players
> t.test(outfield_data$age, goalkeeper_data$age, alternative = "less", var.equal = FALSE, conf.level = .95)

        Welch Two Sample t-test

data:  outfield_data$age and goalkeeper_data$age
t = -3.6102, df = 756.09, p-value = 0.0001631
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
       -Inf -0.4351041
sample estimates:
mean of x mean of y 
 26.22682  27.02690 
```

For age we are performing a Two Sample t-test that the average age of goalkeepers more than average age of outfielders, and we choose 95% confidence interval as it provides a balance between precision and certainty. The null hypothesis here is that there is no significant difference in ages of goalkeepers and outfielders.

The t-statistic measures how many standard deviations the sample means are apart. In this case, the negative value (-3.6102) indicates that the mean age of outfield players is, on average, less than the mean age of goalkeepers. The p-value is the probability of observing such a result (or more extreme) if there is no true difference in the means. A small p-value suggests evidence against the null hypothesis. Here, the p-value is very small, indicating strong evidence that the mean age of outfield players is significantly less than the mean age of goalkeepers. The confidence interval provides a range of plausible values for the true difference in means. In this case, the interval (-Inf, -0.4351041) suggests that, with 95% confidence, the true difference is likely to be less than -0.4351041. The statistical analysis indicates a significant difference in the mean ages between outfield players and goalkeepers. The small p-value provide strong evidence against the null hypothesis, supporting the conclusion that outfield players tend to be younger, on average, than goalkeepers. The 95% confidence interval further emphasizes this difference, suggesting a probable range for the true difference in means.

Conclusion

We were able to examine the dataset and determine its normalcy. We were also able to perform hypothesis testing on a few columns in order to better comprehend t-test statistics and p-values. We were able to interpret under what conditions the null hypothesis is rejected or accepted and use statistical evidence to answer our inquiries. We were able to demonstrate statistically that goalkeepers are statistically taller than outfielders, goalkeepers are statistically heavier than outfielders, and goalkeepers are statistically older than outfielders.

Citations

R Documentation, An introduction to R. Retrieved 15th June 2024 from https://cran.r-project.org/doc/manuals/r-release/R-intro.html#Related-software-and-documentation/

Null hypothesis, Retrieved 15th June 2024 from https://byjus.com/maths/null-hypothesis/