Berke Tüfenk

215140942

CS-240 Project Report

**Section-1:**

1- In all star players data, do players who play more in game, reach higher points?

2- Do players, who get awards, get chosen for all-star team in the same year?

3- Do players, whose coaches get awards, also get awards in the same year?

I’am choosing first question beacuse it is more remarkable for me and I’am really wondering about it. It is also easy to calculate.

I will be focus on relationship between all-star players time duration and their points also how they are related.

**Section-2:**

I will use minutes and points columns from basketball\_player\_allstar database. Due to the fact that i will compare two different measurable data,(time counts with minute,point counts with counting), i will divide into two different time group. First one will be the top minutes which will be including players play more than two period. Other one will be the bottom minutes which will be including players play less than two period. Because of the non values my dataset would be complex, so i replaced empty cells with zeros. Finally, I can work with clear data.

**Section-3:**

For five descriptive statistics we can calculate mean, variance, standard deviation, mode and median.

Mean of points that player scored: 11

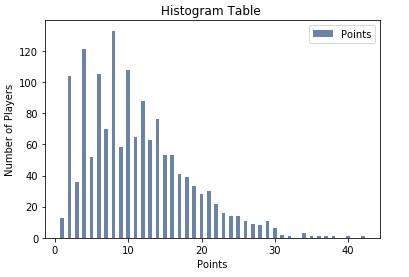
Variance of points that player scored: 46

Standard Deviation of points that player scored: 6

Mode of points that player scored: 8

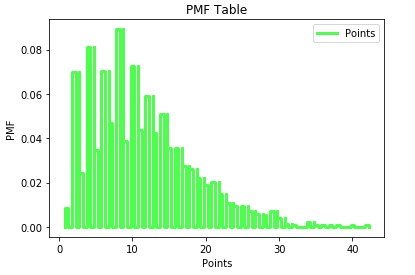
Median of points that player scored: 10

Histogram

****

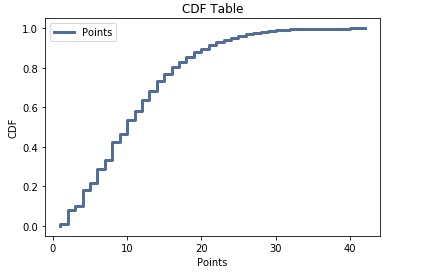
As you see bottom ones are the players that can make less points than the top ones. Furthermore, we can easily see that it is important to find a time to play. Also, we can comment that if players may play to chance more than twenty four minute, he may have a good points avarage.

PMF



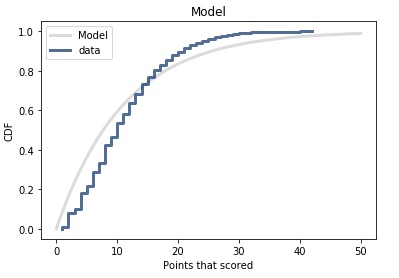
Probability mass function shows that probabilities of possible values. Also, most of points generally scored 4-14 points by all-star players.

CDF



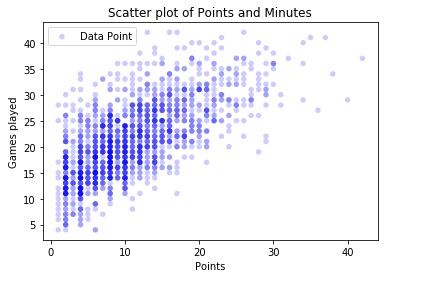
We can see cumulative distribution function, while the highest point is 42, the lowest point is 0 in my data.

**Section-4:**

****

I chosed exponential distribution to represent my data. Although the fact that i tried some different distributions for modelling the data, they did not work as exponential distribution. Generally, we can clearly see that there are two main deviations due to the being a lower from and being a higher from mean points. In the first part of the model, first deviation means all star player who scored higher points than the average, their line seem like lower than the expected. Second deviation will be direct opposite. So, second deviation means all star player who scored who scored lower points than the average, their line seem like higher than the expected.

**Section-5:**

****

As you can see, I used a scatter plot to show the relation of points and all-star players time duration in matches. Also, we can see that both the spearman correlation and the pearson correlation will be positive result. However, there are still outlier points. Moreover, we can see their relation is correlated.

Spearman Correlation: 0.660123398236

Pearson Correlation: 0.664216467917

**Section-6:**

Correlation: 0.664216467917

P Value: 0

I used pearson correlation between time durations and points from all-star players data. P value result is 0. So, my hypothesis will be wrong.

**Section-7:**

To sum up, in this hypothesis , first I decided my question, actually while i was thinking my question, I thought that points and time durations could be correlated with each other. I interpreted the all-star points and I realized some information about all-star players and their points with three steps. First one was histogram of all-stayer players points, second one was probability mass function of all-star players points , final one was cumulative distrubition of all-star player points. I evaluated points and I plot a model for all-star points. This model fitted with exponential distribution. Then, I analyzed my hypothesis according to points and time durations. Finally, Due to the fact that I got the p-value is 0, my hypothesis is statistically significant.