

CS 240

EXPLORATORY DATA ANALISYS



BASEBALL DATA ANALYSIS PROJECT



Ahmed Ahmed Groshar

215382225

Spring 2017

part 1

Apparently from the data available under the appearences excel file its clear that after giving the identification of the year, team, league and player after that the total games value is given followed by 14 distinct posts in which each particular player appeared on the baseball field. Therefore, the performance metric that we are most interested in is the number of appearance in each game by each particular baseball player. Also in the total number of appearances of each player on the field.

I am going to attempt to scrutinize into the relationshp between the two groups of players that batted and players that played on the defence. The two columns 'G\_batting' and 'G\_defense'

Hypothsis: If the appearance of a baseball player as 'G\_batting' is high the appearance of the 'G\_defense' would be high as well.

This means that in other words if a person appears on the baseball field as a 'G\_batting' player that person would be also very competent or would apper on the field as a defense player as well.

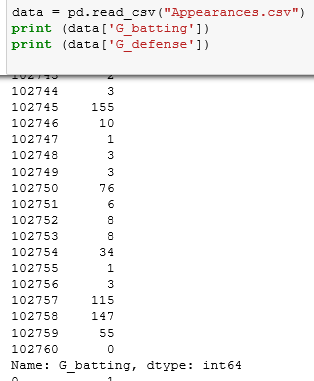
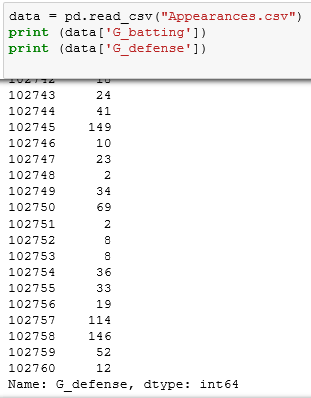
Null hypothesis: If a player does not appear on the field as batting player is it the case that he does not appear as a defense player as well.

Hence, we will attemp to see if there is a significant statistical relationship between these two player’s roles , which would display if a player is a good at batting that very player is very much competent is defense as well.

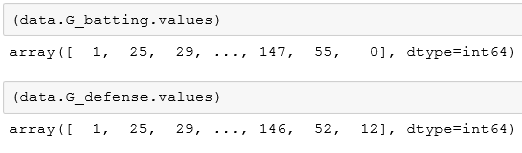
A game of baseball is played between two teams, each composed of nine players, that take turns playing offense (batting and baserunning) and defense (pitching and fielding). A pair of turns, one at bat and one in the field, by each team constitutes an inning. Batting is the act of facing the opposing pitcher and trying to produce offense for one's team. A batter or hitter is a person whose turn it is to face the pitcher. The three main goals of batters are to become a baserunner, drive runners home, or advance runners along the bases for others to drive home, but the techniques and strategies they use to do so vary.



PART2

Here the two columns of interest 'G\_batting' and 'G\_defense' are being read from the csv file and consequently printed as shown above. Since, we have got pretty huge data sets if there is any relation or behavour of the data at certain regions that is related between the columns shall be examined elaborately.



This are the numpy array of the data used for the calculatipon of the p-value.

PART 3

The first graph to be plot is a Histogram of the data in interest. First the Hist module of the thinkstats2 is used then PlePlot of two for the two columns, then it’s plotted and configured.

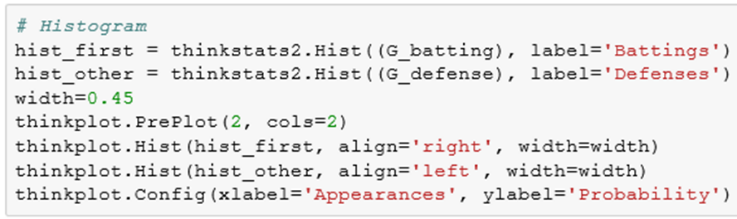
The histogram as a plot lets us to discover, and show, the underlying frequency distribution (shape) of the set close in magnitude two data groups data. This allows the inspection of the data for its underlying distribution. So the distribution of the data in the histogram is defined by the frequency.

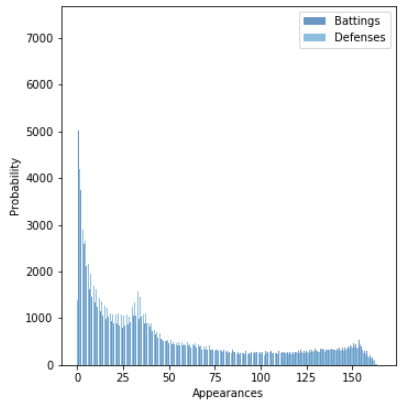
The second graph to be plot is the Pmf which follows exactly the same procedure, however applying the Pmf module.

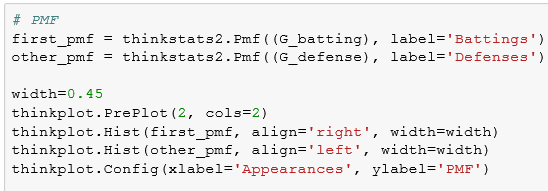
The Probability Mass Function progives us with the probabilities for the discrete random variables. “Random variables” are variables from experiments or in our cose from the data set.

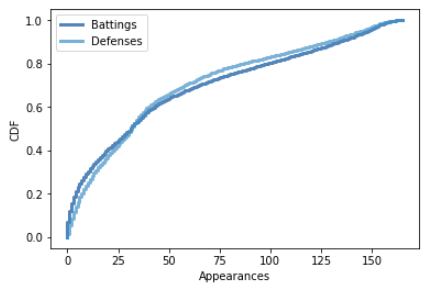
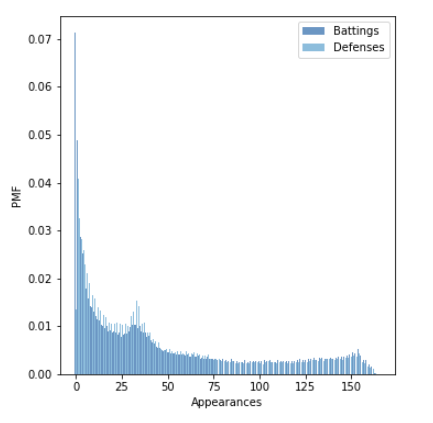
On the third place is the CDF plotted the same way as the others by just using the CDF module. It does give us a clearer view of the picture just by displaying thee very curviture they form .Hence, we observe how similar the data is and how at a single point the domination of the curves change but till keeps the very similar maner of propagation

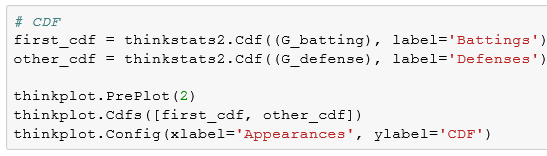
Here their relationship is quite precise and tells a lot about the relationship between the two groups if batting and defence players. Concretly a strong relationship is propagated supporting the proposition that if a player is good at one of the two posts most probably that person would be good in the other too.









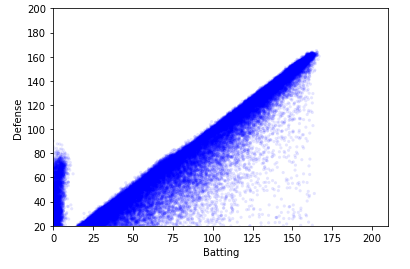
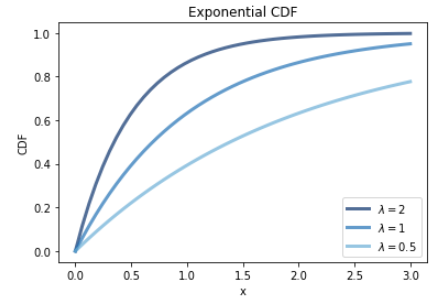
 λ, determines the shape of the distribution.

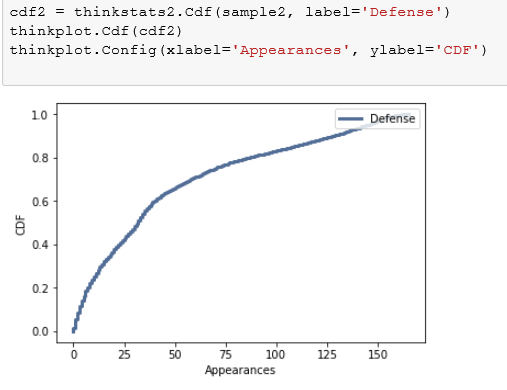
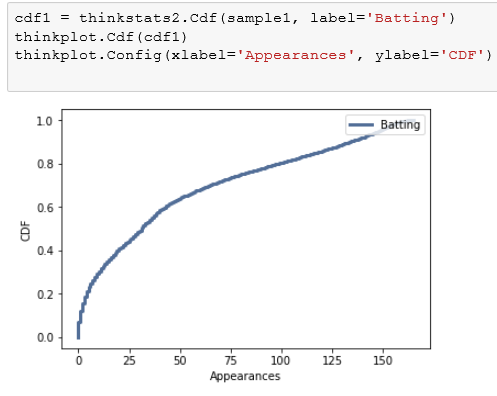
PART 4

In the real world, exponential distributions come up when we look at a series of events and measure the times between events, which are called **interarrival times**. If the events are equally likely to occur at any time, the distribution of interarrival times tends to look like an exponential distribution.

The CDF of the exponential distribution is:

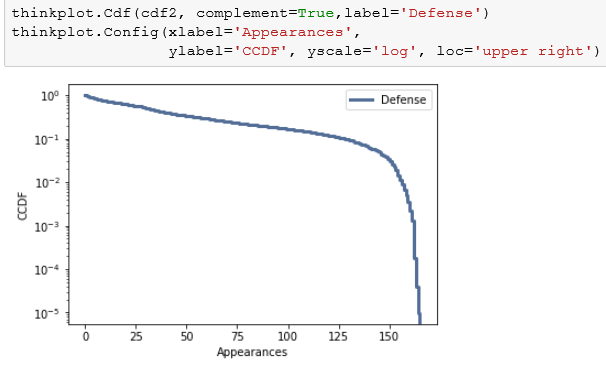
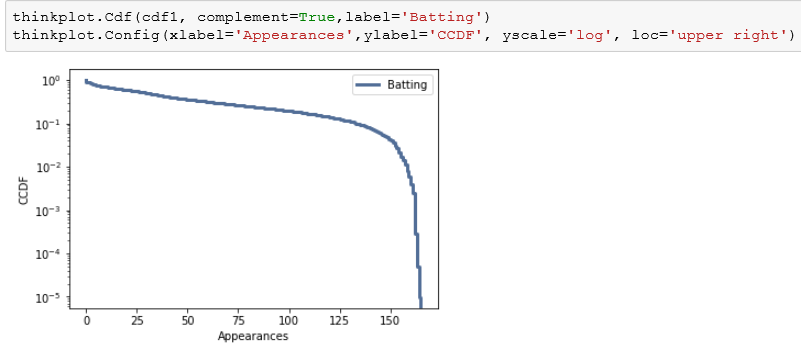
|  |
| --- |
| *CDF*(*x*) = 1 − *e*−λ*x* |

The parameter, λ, determines the shape of the distribution. 



As quite apparent, the shape of the distribution gets closest to the modelling distribution of the exponential distribution function whose λ = 1. The exponential distribution semms to be the onlyone toencompass the data sets from the two different columns. From the individual CDF agoin is easy to see the closeness to their best fit moddeling exponential function.

Also, as visible from the jitter plot the vast majority of the data pairs of the defense and batting performance seems to be very even hence supporting the initial proposition of the hypothsis.



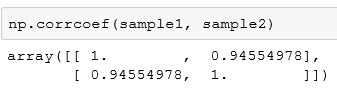
These are the CCDF of the interarrivals on a log-*y* scale. It is not exactly straight, which suggests that the exponential distribution is only an approximation to the real one .

PART 5

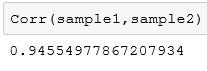
Covariance is useful for some calculations, but it doesn't mean much by itself. The coefficient of correlation is a standardized version of covariance that is easier to interpret.

 Here the samples 1 and 2 are the batting and defense columns. The Covariance shows how strongly corelated two variables are.

In calculating the correlation coefficient the result is a matrix with self-correlations on the diagonal (which are always 1), and cross-correlations on the off-diagonals (which are always symmetric).



Pearson's correlation is not robust in the presence of outliers, and it tends to underestimate the strength of non-linear relationships.The correlation is a single number that describes the degree of relationship between two variables.



Here, the correlation coefficient proves statistically that the their correlation is very high therefore the two groups are vary similar in value aslo in support of the hypothesis.

Spearman's correlation is more robust, and it can handle non-linear relationships as long as they are monotonic.



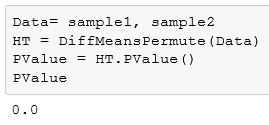
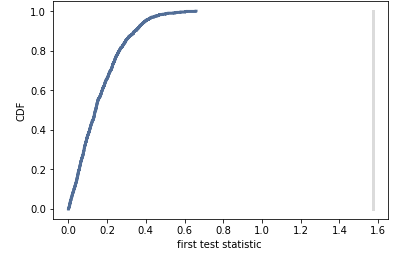
 As seen from the scatterplot above the relationship between your two variables looks monotonic to some extend so by running a Spearman's correlation the strength and direction of this monotonic relationship is measured.

PART 6

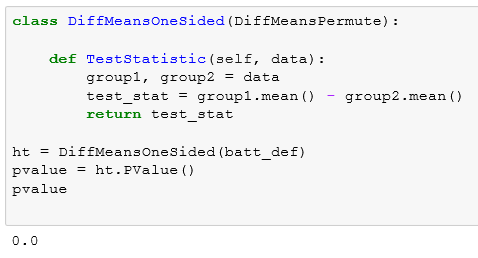
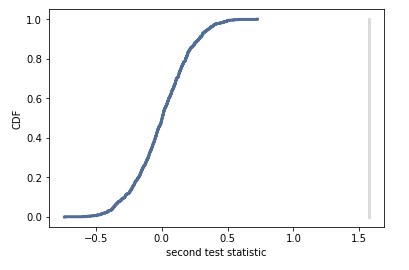
Hypothsis: If the appearance of a baseball player as 'G\_batting' is high the appearance of the 'G\_defense' would be high as well.

This means that in other words if a person appears on the baseball field as a 'G\_batting' player that person would be also very competent or would apper on the field as a defense player as well.

Null hypothesis: If a player does not appear on the field as batting player is it the case that he does not appear as a defense player as well.

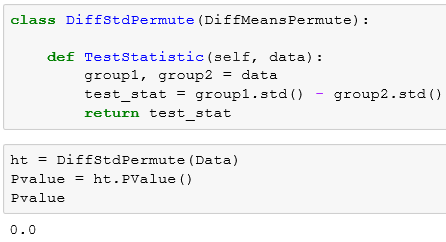
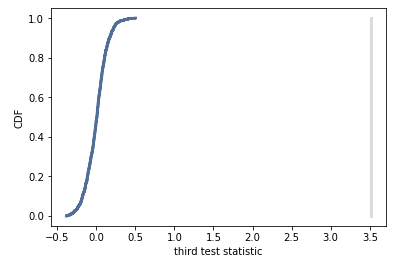


Here, the Test Statistic is difference with the test\_stat = abs(group1.mean() - group2.mean()) which the difference between the group’s mean values followed by the absulute value taken so that a positiva value is obtained at the and becouse the test statistic will return a single number. The value of the p value is very small approximated to zero which means that the p value is significant and therefore their is high confidence in discarting the null hypothesis and supporting the proposed hypothesis to be true. Also, from the graph its apparent that the manner of the curviture is exponentially propagating and we observe how far the p value is resulting in a 0.0 value.



Here, the Test Statistic is DiffMeansOne Sided which uses

test\_stat = group1.mean() - group2.mean() which takes the mean of the two column of interest G\_batting and G\_defense and subtracts the second from the fisrt one differing with the absence of the absolute value which model the data more as a whole becouse negative value may affect the final value of the test statistic. If there are great differences between the data by using this test statistic we would be able to detect it to some extend. Again the here as well the p value is in strong support of the hypohesis by poiting to the null hypothesis as wrong . Negeting the null hypothesis means that the p value is statistically significant and there is indeed relationship between the two groups as stated in the hypothesis. In here, the curviture is a bit different in the base from the previous CDF however, the p value is a bit closer but not close enought to two significant figures.

  
 In the third try of this experiment another test statistic is implemented using

**test\_stat = group1.std() - group2.std()** which utilizes standart diviation statistic on the data groups. Standard deviation is a measure of the dispersion of a set of data from its mean. If the data points are further from the mean, there is higher deviation within the data set. Standard deviation is calculated as the square root of variance by determining the variation between each data point relative to the mean.

As obvious from the p value the significance is confirmed once again and that is quite clear from the CDF plot as well where p value even is even further confirming the preceeding observations.

PART 7

Conclusion:

All in all, the aim of the invstigation in this report was two prove that there is a significant relation between the two sets of data provided under the namings G\_batting and G\_defense. From the various kinds of plots an observation was conducted that concludes from the emphirical evidence from the proof of the hypothsis analysis that the proposed initially hypothesis is correct. Hence we have witnessed an interesting investigation that showed that if a player is good at batting that very player player is as that much good at defence on the baseball field. By using different test statistics we were able to show a very strong value for the p-value which consequently directs us to the trithness of the hypothsis. The valuous distribution plot anabled us to observe the distribution of the da two sets of data which additionally helped us imagine how the data is scatered on the the plane it accupies. The strong support from all the statistical methods used leads us to conclude that the statistical significance of the observation is present both visually from the alternating kinds of plots and numerically from the statistics applied.