**CS 240 PROJECT**

**Ubeydullah Osman Tamer**

**214001771**

**Introduction:**

In this project, we are going to analyse baseball statistic data from 1871 to 2016. We will determine 3 different questions about this baseball statistics. Then we are going to show our works on one of them.

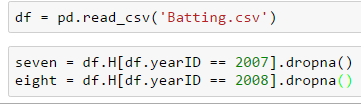
**Part 1:**

In my project, I determine 3 different questions

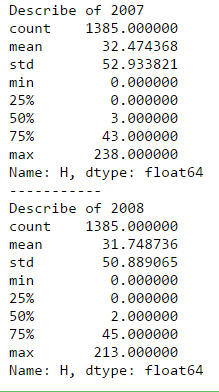
* Is there any relationship between the performances of baseball players with previous year? If a baseball player end up a season with a good hit rate, next year can he catch the same rate?
* Is there any relationship between the salaries and the type of the league? National League players earn more than American League players?
* If manager appear in the match, the probability of the win that team is increase or not?

I choose to analyse first question. My hypothesis is “There is a strong correlation between the hit rate in 2007 and 2008”.

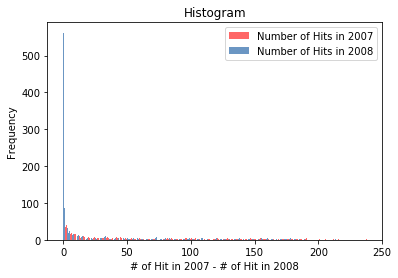
**Part 2:**

First of all, I define my data frame that I am going to use. I will use “Batting.csv” file. Then I defined the columns that I am going to use. Since I will compare the hit rates in 2007 and 2008, I create two variable called seven and eight.

**Part 3:**



As you can see, I printed out some relevant statistics. Their mean and standard deviations are very close to each other. In 2007, a baseball player hit the ball almost 32 times on average, in 2008 its almost 31. The standard deviation of hits in 2007 is almost 53. Also in 2008, standard deviation is almost 51. Since standard deviations are pretty high, which means number of hits distributed widely. Variance is also give us information about the distribution.



As you can see, this figure shows you the number of player frequency by number of hits in 2007 and 2008 between 0 and 100 hits. If you analyse deeper, as number of hits in 2007 decrease, in 2008 number of hits may increase or decrease. Frist I used the .Hist() function to find the frequencies of data. Then I used .PrePlot(2) function to draw to one graph but two values. Then I .Show the graph.

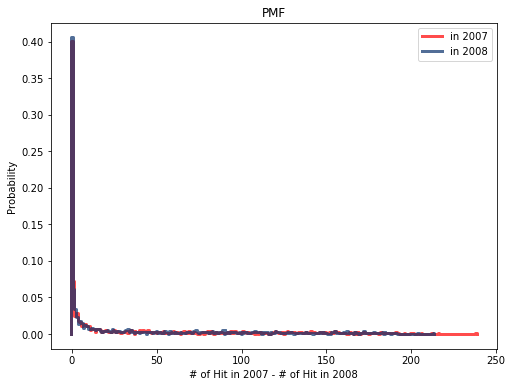
**Central Point:** Most of the player cluster around 0 hit.

**Modes:** Most of the player between 0 hit and 20 hits.

**Spread:** In 2007 max number of hits is 238, in 2008 its 213. It’s difficult to realise on this figure, but we can see it from .describe() function.

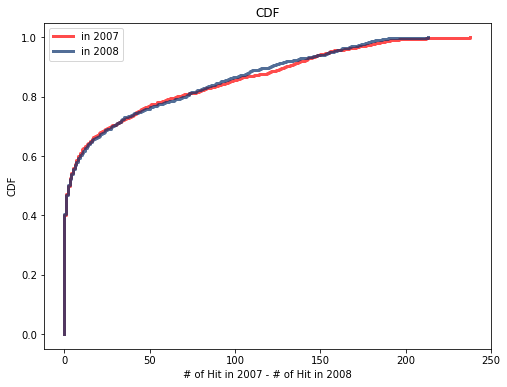
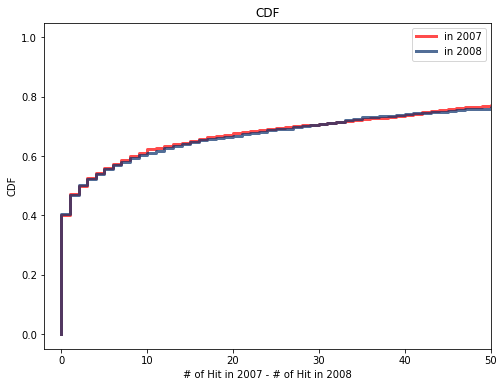
**Tails:** As the number of hits increase but, the number of player mostly decrease.

**Outliers:** There is no outliers.



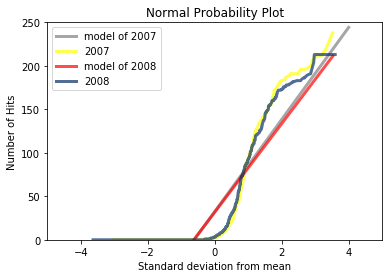
The I used .Pmf() function to find probability of each number of hits. Then I draw it on graph with using the same steps on histogram graph. In this graph it’s really difficult to compare the values between the lines. We can also use PercentileRank() function to compare a score with one of the data. You can check the codes. That function can show you, where are you in the distribution. For example, if you hit the ball 50 time, you are better than 77% of the players.

To read CDF graph, we need to look up percentiles. But the figure on the left cannot give me enough details. So I used xlim=[0,50] parameter to see the from 0 hit to 50 hits. As you can see the figure on the right, 40% of the players couldn’t hit the ball in both year 2007 and 2008. Also 60% of the players hits the ball less than 13 times in both year 2007 and 2008.



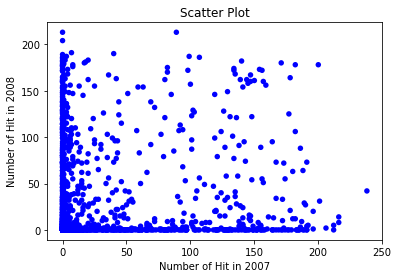
**Part 4:**

This figure shows the normal probability distribution. Firstly we create two model for our data with using mean and standard deviation. We can see that until players reach the 75 hits, they are under the model, after 75 hits, they are over the model. Each time we try to plot the graph, it will change but only the tails. Other parts are the same. Since most of the people hit 0 time, distribution is zero at the head until standard deviation from mean approach to 0. In both 2007 and 2008, distribution didn’t fit with the models. For both years, model is almost the same, but distributions are different and they never follows the models. So the distribution is not normal.



**Part 5:**

In this part we are going to analyse the correlation between the hit rate in 2007 and 2008. Firstly, I plot the Scatter Plot graph with Scatter() function. Surprisingly, there isn’t a correlation between the player’s hit rate in 2007 and 2008. Most of the players gathered round 0 hit. We may say that if a player has a good hit rate in 2007, he couldn’t hit the ball next year. Since we don’t have a very big data no need to jittering.



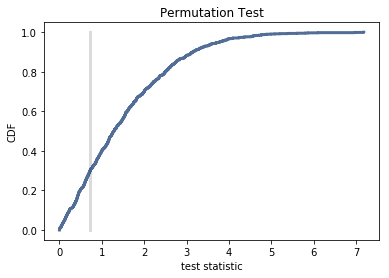
 Now we are thinking that there is no correlation between hit rates in 2007 and 2008 because some players never hit the ball in 2007 but in 2008, hit more than 100. But we need more evidence. I used Corr() function to do that. This function’s output is telling the strength of relationship between two variables. Also I multiply the correlation with 100 so we can see well. The correlation between hit rates in 2007 and 2008 is 5.4% which is really low. So we can say that there are almost no correlation between these two variables. Also Cov() function measure the tendency of two variables to vary together. Covariance is useful for some calculations but it doesn’t mean much by itself.

**Part 6:**

My hypothesis is “There is a strong correlation between the hit rate in 2007 and 2008”. But till now, we couldn’t find any correlation between these two variables. My null hypothesis is “There is no correlation between the hit rate in 2007 and 2008”.

Now I am going to test my hypothesis if the effect is statistically significant or not. To be able to say that, we need to find p-value. If p-value is low the effect is statistically significant, which means it’s unlikely to have occurred by chance. In that case we infer that the effect is more likely to appear in the larger population.

The test statistic is 0.72 which is the absolute difference of the means. A natural choice for the test statistic is the difference in means between the two groups. I calculated the p-value by using DiffMeansPermute() class which also use HypothesisTest() class from thinkstats2. The test statistic is the absolute difference in the means. First we combine the groups into one NumPy array. Then RunModel() function simulates the null hypothesis by shuffling the pooled values and splitting them into two groups with size length of the values. To test the difference in performance I find the values of my data. Then use DiffMeansPermute() function and I calculate the pvalue with .PValue() function. The p-value of my data is 71%. Since the p-value is too high, which mean this effect occurred by chance. So this effect might not be generally true in the population.



**Part 7:**

At first my hypothesis was, “There is a strong correlation between the hit rates in 2007 and 2008”, but we couldn’t observe that. Most of the players don’t have unsteady. If they had a good hit rate in 2007, mostly they are not able to catch the same rate as well. Also we calculate the p-value as 71% which is really high. P-value shows that this values are not statistically significant. So I cannot validate my hypothesis.