EXPLORATORY DATA ANALYSIS

(CS 240)

PROJECT

Sümeyye ÇİLDAN

213952662

Part 1 : Brainstorm

My Question:

What is the relationship between W and HRA? Are there any strong relationship between them?

My Hypothesis:

If homeruns allowed decrease, number of wins increase.

Part 2 : Data Analysis

I used Table.csv of Baseball Data. I chose the W and HRA columns. W represents the wins and HRA mean is homeruns allowed.

I used pandas and read_csv for read the csv file.

```
data = pandas.read_csv('Teams.csv') #read database

w = data.W # Wins

hra = data.HRA # Runs
```

Part 3 : Histogram , PMF, CDF

I used min(), max(), mean(), var(), and std() functions and I defined these values.

For wins max value is 116, and for homeruns allowed. Both of them have 0 for min value.

```
(0, 116, 0, 258)
```

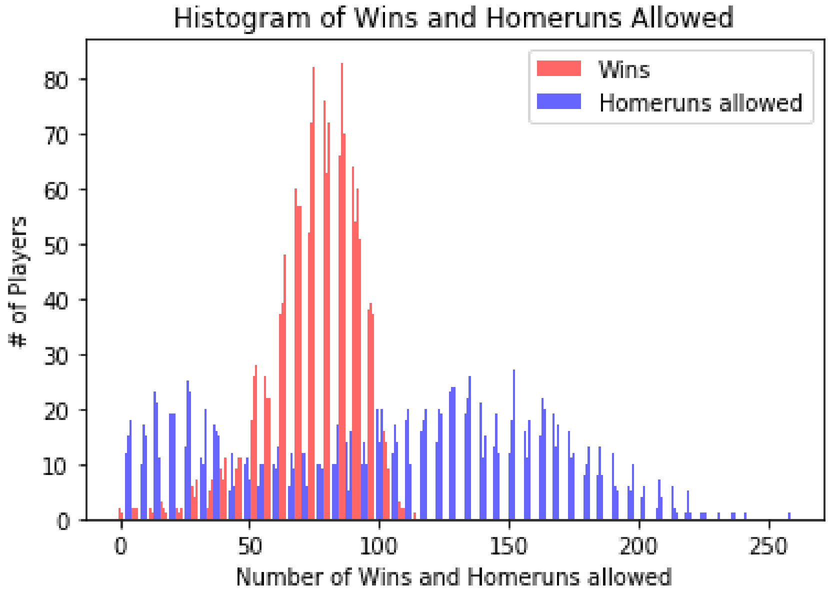
Mean of HRA is grater than Mean of W.

HRA variance is 3439.67 and it is approxiametly 10 times of W variance.

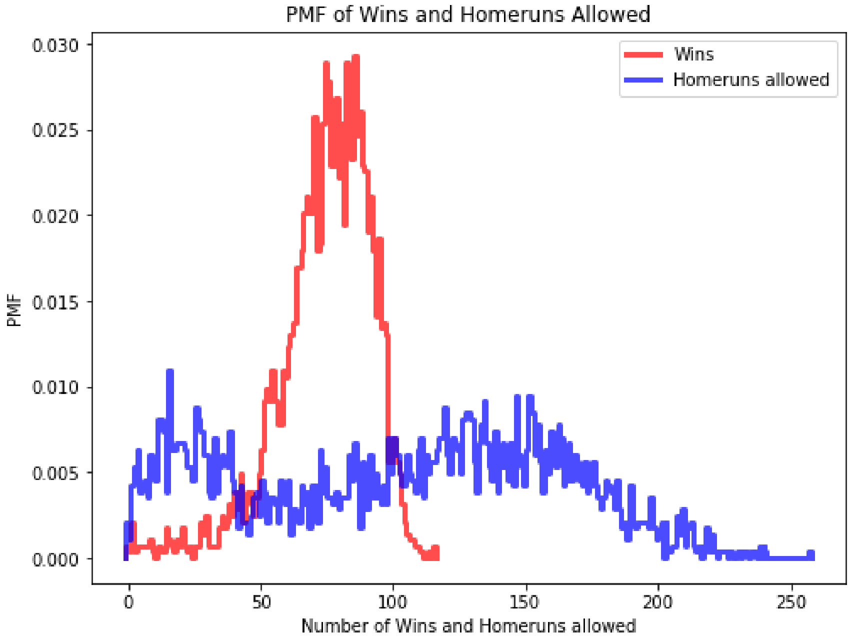
Std of W is less than std of HRA. HRA value is 3 times of W value.

In short, all homeruns allowed values are more than wins values.

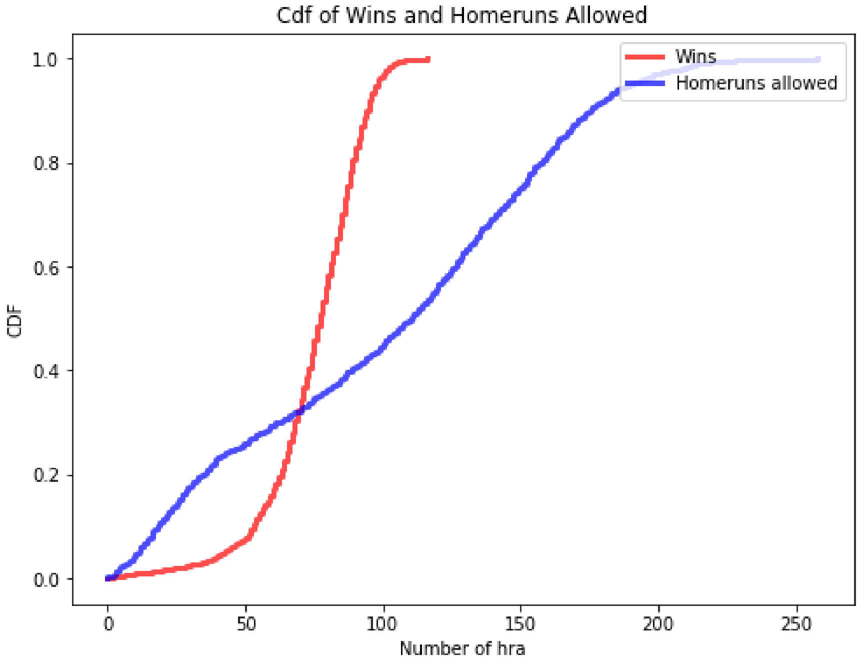
```
Mean of W: 74.81410934744268
Variance of W: 309.45061242981785
Standard Deviation of W: 17.591208384582846

Mean of HRA: 102.04514991181658
Variance of HRA: 3439.676859848728
Standard Deviation of HRA:58.64875838283985
```



This histogram shows number of wins and homeruns around. X-axis represent the number of wins and homeruns allowed.



In PMF, there are probabilites of wins and homeruns allowed. For both of them, probabilities are flactuated. However, for homeruns allowed values are closer, and max value is about 0.011. When homeruns value is maximum, its probability is minimum. Wins probability increase until 80, and it has maximum probability in this point, after that it decrease and in approximately 120, probability is minimum.

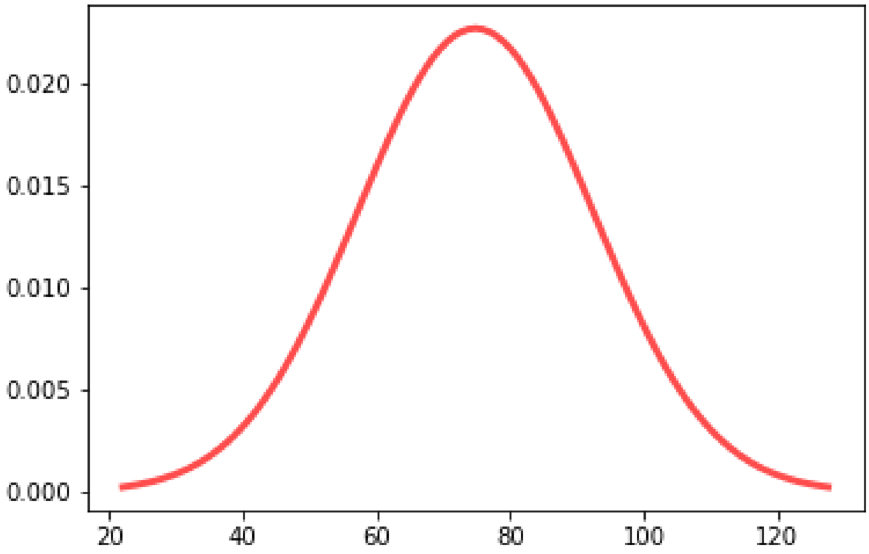


In CDF, both of them incrase. Wins value is increase faster.

PART 4 : Modelling Distribution

I used normal pdf distribution.

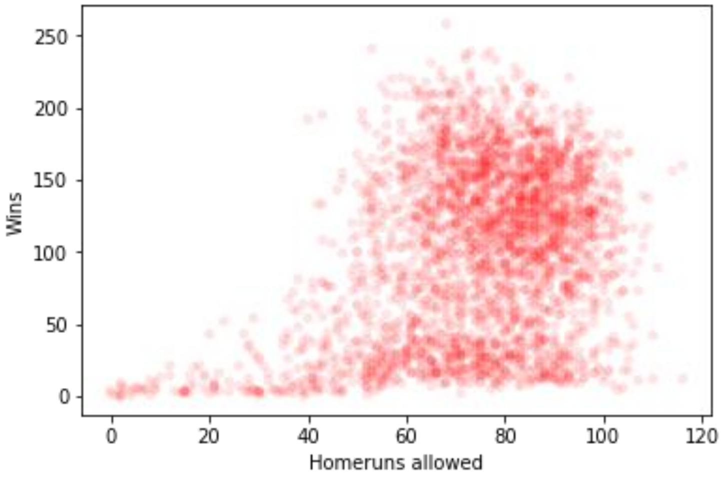
'Normal Pdf Distribution of Wins'



PART 5 : Correlation

```
[[ 1.          0.31970224]
 [ 0.31970224  1.        ]]
```

I controlled the numpy's correlation coefficient of wins and homeruns allowed. I obtain 0.31970224 value, so correlation value is low.



PART 6 : Hypothesis Testing

I used hypothesis testing and applied 4 steps. Firstly, I chose test statistic, later I defined the null hypotesis. Next step, I compute the p-value, finally I interpreted the result.

I used thinkstats2.HypothesisTest in class. I chose test statistic and compare wins and homeruns allowed.

Test Statistic : There is a relationship between wins and homeruns allowed. If homeruns allowed decrease, number of wins increase.

Null Hypothesis : There is a no relationship between wins and homeruns allowed.

`p-value: 0.0` . It is statistically significant.

PART 7 : Conclusion

In conclusion, there is similarity between wins and homeruns allowed. When I checked the correlation, similarity between them was low. In hypothesis testing, test statistic was statistically significant. Therefore, there is a relationship between wins and homeruns allowed.