Layla Idris

215362797

CS 240- Exploratory Data Analysis

Final Project

**Introduction:**

We have a baseball statistics from 1871 to 2016, we will analyze the data and use some modules to make a Hypothesis test and to make a data frame and put the data in it and plot them too.

**Part 1:**

We will ask a specific question about a chosen file to analyze multiple stuff.

“Managers” file is used and the question is: can we analyze the relationship with winning and losing? If they win, does it mean that the total of losing got less? Is there a positive relationship or negative relationship?

**Part 2:**

I am going to use two columns from the selected file, they are the Winning “W” and the Losing “L” file. I will observe the relationship between them if it is found. So after reading the file I printed the columns.

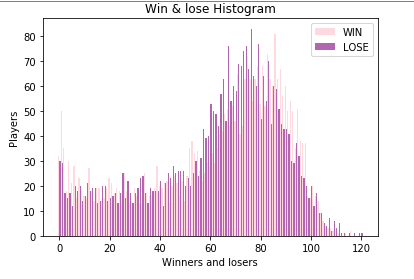
**PART 3:**

I looked on the columns in a more detailed way and found the mean, maximum, minimum, standard deviation and the variance.

I used the built in functions to find them.

We found that the mean, the maximum and the minimum are almost the same and the standard deviation and variance differ a bit.

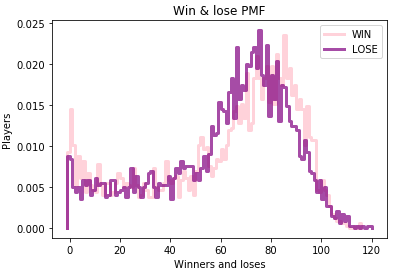
I made a histogram of the data collected from the two columns, I used “thinkstats2.Hist()” and selected the color and label names as I wanted

And to plot it I used “thinkplot.Hist ()

This Histogram shows how many players won and lost

I made a Pmf of the data collected from the two columns, I used “thinkstats2.Pmf()” and selected the color and label names as I wanted

And to plot it I used “thinkplot.Pmf ()”

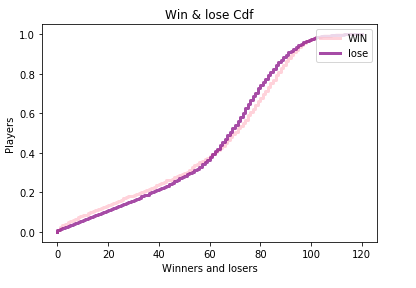


This Probability of Mass Function (PMF) shows how many players won and lost

I made a Cdf of the data collected from the two columns, I used “thinkstats2.Cdf()” and selected the color and label names as I wanted

And to plot it I used “thinkplot.Cdf()”

At the end I used “thinkplot.Show ()” so that the plots will show up.



This Cumulative Distribution Function (CDF) shows how many players won and lost

**Part 4:**

I used a Normal Probability of Mass Function distribution so that I can see the density of my distribution at a very specified location on the standared deviation according to the mean. That’s why I found the mean at the first and I estimated KDE. KDE: is an algorithm that estimated a PDF based on a sample.

After finding the mean I used “.NormalPdf” and then according to the inserted mean and the inserted standard deviation I created a normal Pdf . This was for the winning part and then I did the same thing for the losing part and at the end I called the pdf.

Then I found the median of the losing and winning columns by using “Median ()”.

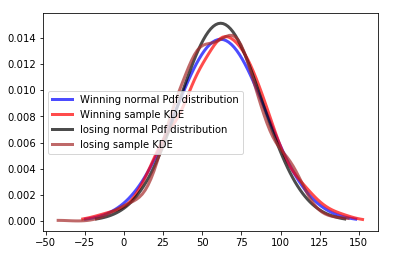
After that I used the density function to find the density of the winning and losing parts which is “.Density ()”

I did draw a normal distribution of them and I used “thinkplot.Pdf” to do that and I did choose different colors to compare them and named the labels.

I wasn’t able to use KDE because I didn’t randomomize all the means that I have and the stardard deviation. So I used “np.random.normal ()” and I put 1000 at the randomize part and 1000 for the pdf to be iterated 1000 times.

At the end I plotted everything and showed it.

This is my result:



**Part 5:**

We found the covariance and the correlation in this part so that we will be able to answer our questions the covariance will show us the tendency. The correlation will show us the strenght of how much relationship there is between the two variables.

By using “np.asarray” we convert into Numpy arrays so that we change the xs and the ys from the python sequence. After that we use Cov function to compute from the sample mean so I used winning and losing as my parameters.

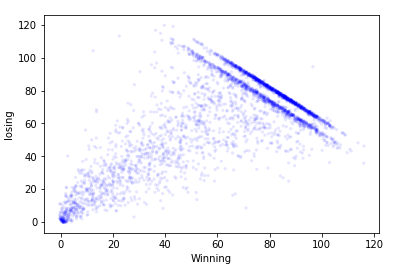
My covariance is very high that means that the assets move together.

Then we found the correlation by using Pearson’s correlation, if the answer is positive then the correlations is positive that means that if one variable is high the other is high too. However, when it is negative it means that when one variable is high the other is low. And my correlation is positive.

We found the percentage of the correlation and the covariance by multiplying it with 100.

After that I plotted by using “thinkplot. Scatter ()” and used a high and a weight and a jitter of 0.3

This is the result:

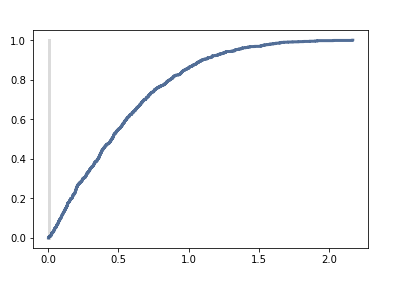


In the parts that have a dark color it is shown that losing and winning have a big similarity.

**Part 6:**

I used a bultin function which is “HypothesisTest” and applied it to test and check my hypothesis.

If there is a high relationship between winning a game and getting the total of losing get less then it is called Test Statistics. However, if there was no relationship then it is a Null Hypothesis.

****

P-value = 0.994 so there is a realtionship.

**Part 7:**

Taking everything into consideration, there is a big similarity between winning and lossing. I showed that relationship in my histogram, pmd and cdf. Furtheremore, I checked the covariance and the correlation. And as an end I ran a Hypothesis Test and showed that the relationship is high and I answered my question,

There is high positive relationship between winning and losing.

**References:**

The book: Think Stats**.**