Miles A. Vendetti-Houser

ECE 3522: Stochastics

Department of Electrical and Computer Engineering, Temple University, Philadelphia, PA 1912

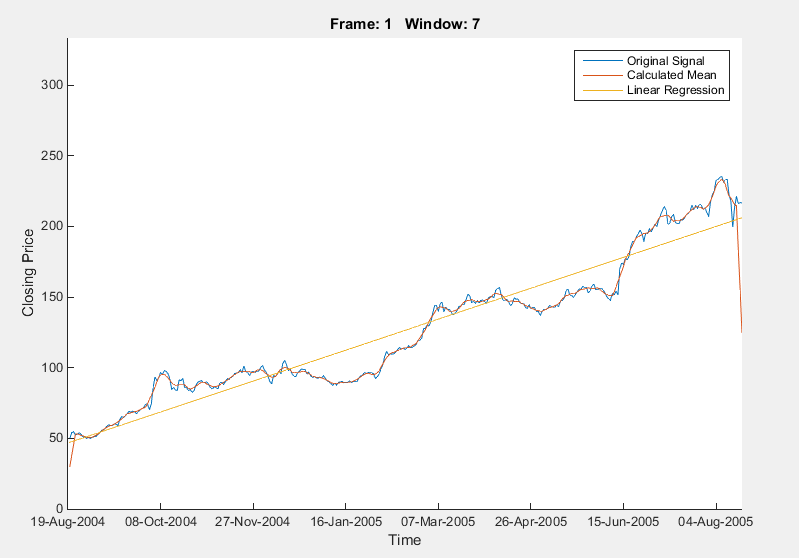
# Problem Statement

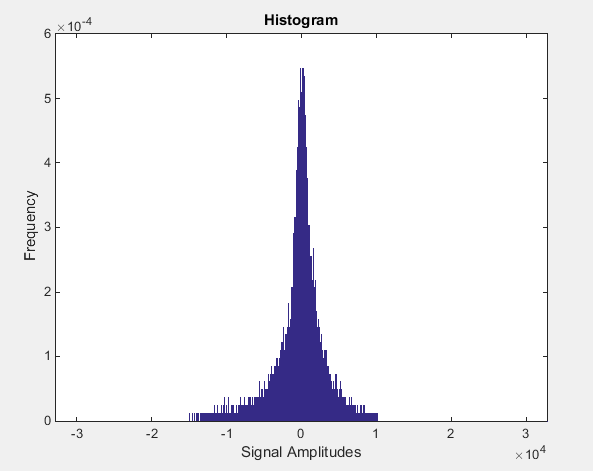
The assignment is designed to familiarize students with primitive predictive analytic techniques. Where according to Wikipedia, “*Predictive analytics is an area of data mining that deals with extracting information from data and using it to predict trends and behavior patterns*.” One would assume that if we are using the data to predict trends that this area of statistics must be very popular and implemented in many fields. In fact, these techniques are used throughout many fields such as risk management, machine learning techniques, and neural networks just to name a few.

# Approach and Results

The first step for proceeding with the assignment was to import the data from Google’s stock prices since its inception and extract the necessary information such as the closing values and corresponding dates for the values. Then, using a frame size of one day, and window of a week, we computed the arithmetic mean or average on each iteration of a frame. The motive for computing the mean is to represent a number of measurements or values as a single value. The frame and window size chosen reveals the closing price average per week. The mean can provide insight regarding trends in stock fluctuations on a week by week basis. The mean for discrete random signals is computed using the following equation:

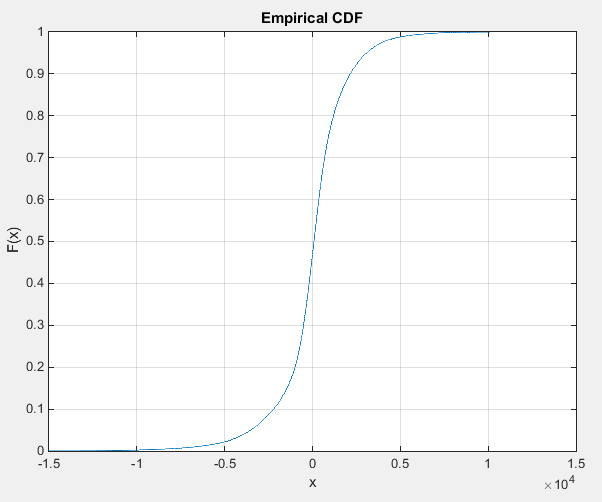
Where n in this case represents the windows size. The resulting values over Google’s stock file was plotted over the original prices. Next, we computed a linear regression using the data. The linear regression slope should be thought as ‘the expected value of Y for a given value of X’. The linear regression line plotted over the mean and original signal implies that the closing prices are increasing. Basically this is saying that if you were to invest stock in Google, over time, your stock will increase. The linear regression follows the overall trend of the computed mean in that its value (slope) is positive (increasing). The plot for part one is shown below.



Part two, the approach for this part required some more time consumption outside of MATLAB and more so researching histograms and there usage. Not because of the intensity of the assignment, but more to become familiar with histograms and what they tell us. After a couple hours of research and practice building histograms by hand (beginning with bins of size 1 and increasing bin sizes in order to master the concept), it all began to make sense. Also, histograms remind me of plotting in the Signals course, where we would plot amplitude versus frequency except…inverted? Anyway, I used the histcount()function in MATLAB to plot the figure below.

I created variables that comply with requirements for the assignment. For example, numbins was defined as the entire data length divided by the desired binwidth of 10. The functions allows the change of specific ‘*name-value*’ pair attributes. According the MATLAB documentation, the Normalization attribute can be changed from the default ‘count’ to ‘probability’. Using ‘probability’ as the newly specified value makes each N value equal to the relative number of observations in the bin such the sum(N) is 1. If I am not mistaken, that sounds like normalizing each bin over the entire data set. If I am completely wrong in choosing this, then I will argue that over the course of completing this assignment, the overall shape of the histogram did not change despite the value of f(x).

The plot below displays the cumulative distributive function using the same information for the histogram shown above.



The above figure portrays the integral of the probability density function. By visual inspection, one may point out that this plot also looks like the area of the histogram we previously plotted. The area converges to 1 as x goes to infinity. This is an example of a property of the PDF which satisfies the equation;

Where fX is called the probability density function.

# MATLAB Code

%Part One

clear**;**clc**;**close all**;**

addpath **../**01

%variable declarations

filename **=** 'google\_v00.xlsx'**;**

audio\_file**.**location **=** 'rec\_01\_speech.raw'**;**

YEAR **=** 365**;**

%import the data

**[~,~,**raw**]** **=** xlsread**(**filename**);**

% import the audio signal

ff1 **=** fopen**(**audio\_file**.**location**,** 'r'**);**

audio\_file**.**data **=** fread**(**ff1**,** inf**,** 'int16'**);**

fclose**(**ff1**);**

% Seems to run differently on my Mac OS and Windows PC, so let's fix that

% if PC run the following lines...

**if(**ispc**)**

%convert date string to numeric type

formatIn **=** 'mm/dd/yyyy'**;**

%put data into allocated vectors

data**.**date **=** datenum**(**char**(**raw**(**2**:**YEAR**,**1**)),** formatIn**);**

% otherwise, run this line

**elseif(**isunix**)**

data**.**date **=** cell2mat**(**raw**(**2**:**YEAR**,**1**));**

**end**

data**.**close **=** cell2mat**(**raw**(**2**:**YEAR**,**5**));**

% Window / Frame Analysis

M **=** **[** 1 **];**

N **=** **[** 7 **];**

full\_mean **=** zeros**(**length**(**M**),** length**(**N**),** length**(**data**.**close**));**

% loop over the a set of frame/window combinations.

**for** m **=** 1**:**length**(**M**)**

% set up a plotting window and label it

%

h1 **=** figure**(**'name'**,** 'Google Stock Closing Values'**,** 'numbertitle'**,** 'off'**);**

**for** n **=** 1**:**length**(**N**)**

% call a function to compute the rms vector

%

full\_mean**(**m**,**n**,:)** **=** compute\_mean**(**data**.**close**,** M**(**m**),** N**(**n**));**

% label the plot:

% include information about the parameters for each plot

%

figure(h1);

str = sprintf('Frame: %d Window: %d', M(m), N(n));

% plot the rms contour

%

t = linspace(1,length(full\_mean), length(full\_mean));

[a0,a1] = Linear\_Regression(t, data.close');

y = a0+a1.\*t;

hold on;

plot(data.close,'DisplayName','Original Signal');

plot(squeeze(full\_mean(m,n,:)), 'DisplayName', 'Calculated Mean');

plot(t, y', 'DisplayName', 'Linear Regression');

hold off

n = datenum(2004,08,19);

dateaxis('x', 1, n)

axis([0 length(full\_mean) 0 (max(full\_mean)+100)])

legend('-DynamicLegend')

%legend('Closing Prices', 'Mean Value','Linear Regression')

% label the axes

%

title(str);

xlabel('Time');

ylabel('Closing Price');

end

end

%Part two

clear**;**clc**;**clf**;**close all**;**

addpath **../**01

filename **=** 'rec\_01\_speech.raw'**;**

ff1 **=** fopen**(**filename**);**

sig **=** fread**(**ff1**,** inf**,** 'int16'**);**

fclose**(**ff1**);**

binwidth **=** 10**;**

norm **=** 'probability'**;**

numBins **=** round**(**length**(**sig**)** **/** binwidth**);**

binlimits **=** **[-**32767**,** 32767**];**

**[**N**,** edges**,** bins**]** **=** histcounts**(**sig**,** numBins**,** 'BinLimits'**,** binlimits**,** ...

'Normalization'**,** norm**,** 'BinWidth'**,** binwidth**,** 'BinMethod' **,** ...

'integers'**);**

bar**(**edges**(:,**1**:end-**1**),**N**)**

axis**([**binlimits**(**1**)** binlimits**(**2**)** 0 6**\***10**^-**4**])**

xlabel**(**'Signal Amplitudes'**)**

ylabel**(**'Frequency'**)**

title**(**'Histogram'**)**

# Conclusions

The histogram plot for the speech signal displays the frequencies of amplitudes from -32767 to 32767. The histogram tells us that due to its asymptotic property towards x = 0, the signal crosses over the x-axis more than it remains in either positive or negative ranges. Some information I gathered from the assignment is if you notice the negative and positive corresponding amplitudes both contain similar frequencies. If the only computation we were told to do on the audio file was to create a histogram, we would still be able to conclude key details from the plot. One detail is positive x-values mirror negative x-values in terms of the frequency. This implies that there is no offset on the original signal. Also, the low frequencies towards the edges of the plot tell us that there is not a whole lot of talking in the signal. The frequencies corresponding to silence or small levels of noise outweigh the frequencies relating to amplitudes recorded from say, someone’s voice, or any large value of noise.