Joe DeMarco CA04

The red line shows a normal pdf. The blue graph is a histogram of the voice data. This did not look right to me because I normalized the histogram data. As a check I added the normalized histogram values up and they do indeed equal 1. Also I know the normal pdf by definition should equal 1. But if I had to guess I would say the blue area is 3 or 4 times the size of the red.



I think this is because there are unseen empty bins in the blue data which are spreading it out. I am trying to reduce this effect by making fewer bins. The first plot has 82425 bins, the second plot has 10000 bins. The hypothesis seems to hold but I’m not sure how to correct this error completely.



And here I zoomed in super close on the 10000 bin plot and it still looks like there are a lot of empty bins



So the normal plot looks like it is an ok fit for the voice data. Though the true distribution is probably much more narrow than normal, I didn’t see a distribution that fit better.

This is a plot of the google data. The blue is a histogram plot, the red is a normal distribution, magenta is exponential, cyan is a Rayleigh distribution.



The google data also has the same problem as the voice data in that it has a bunch of empty bins. I again tried to compensate for this by using a larger bin size. I settled for a bin size that is a factor of 10 lower than the total number of samples.

It is apparent that the normal distribution is a terrible fit for this data set, so I tried a couple of other distributions. The exponential fits better, but the actual data is not as clumped around zero as the exponential needs it to be.

The Rayleigh distribution seems to fit the data reasonably well. The peak lines up and the values trail off on the edges in a similar way.

If I am interpreting these plots correctly the Rayleigh distribution would seem to say that the next most likely value for the stock price to be would be ~250. But we know this is not accurate, the actual stock price the day after this data ends was ~450.

It seems like the results will vary widely depending on what bin size I choose, and what data I decide to use. If I were to choose a different bin size, or choose to use less of the data set (ex. Only the last few months) I may find that totally different distributions fit the data.