ISYE 6501, Week 3 HW

**Question 1**

Using crime data from http://www.statsci.org/data/general/uscrime.txt (description at http://www.statsci.org/data/general/uscrime.html), test to see whether there is an outlier in the last column (number of crimes per 100,000 people). Is the lowest-crime city an outlier? Is the highest-crime city an outlier? Use the grubbs.test function in the outliers package in R.

**Response –**

Upon testing with shapiro.test in R, the p-value is 0.001882 which says that the null hypothesis of data being normally distributed is to be rejected. But upon plotting the data in qqnorm, it can be seen that the data is almost normal and the low p-value might have been caused by the outliers in the data. Hence, we can use the grubbs.test function in R.

Grubbs.test, type 11 – p-value is 1 which means the test failed to reject the null hypothesis that there are no outliers in both tails. Hence, there might be outliers present or they might not be.

Grubbs.test, type 10 – p value is 0.07. The outcome of this test depends on the acceptance threshold for the test. If the threshold is 0.1, null hypothesis can be rejected but if the threshold is 0.05 the test failed to reject the null hypothesis. Observing the plotted data in qqnorm test, it can be said that the data point with value 1963 might be skewing the data. Hence Grubbs.test type 10 is again performed without this data point, and this time the p value is 0.02 which rejects the null hypothesis that there is no outlier in one tail. This indicates the highest value 1993 is an outlier. On the opposite tail, the p-value is again 1, hence the test failed again to reject the null hypothesis that there is no outlier in the tail of smaller values.

It can be said with certainty that the highest value 1993 is an outlier, if we don’t consider the datapoint “1963” or have a threshold value of 0.1.

**Question 2**

Describe a situation or problem from your job, everyday life, current events, etc., for which a Change Detection model would be appropriate. Applying the CUSUM technique, how would you choose the critical value and the threshold?

**Response –**

A change detection model can be used to distinguish between the products with defects and without defects in a manufacturing process. The parameters could be dimensions of the product or the density of the product. This model could help in identifying the small shifts in the average from the process target dimension/weight tolerances. Once the change or deviation is detected, then the manufacturing process could be paused to fix the issue and keeping the downtime to a minimum.

The critical value and the threshold value will be chosen based on the cost impact the shutdown of the manufacturing process causes. For example, if the shutdown is more costlier than the cost of poor quality of the products then the critical value and threshold will be set to capture the change after significant change occurs. Whereas, if the cost of poor quality is much high then the critical value and threshold will be set to capture the defect very early.

**Question 3**

1. Using July through October daily-high-temperature data for Atlanta for 1996 through 2015, use a CUSUM approach to identify when unofficial summer ends (i.e., when the weather starts cooling off) each year. That involves finding a good critical value and threshold to use across all years. You can get the data that you need from the file temps.txt or online, for example at http://www.iweathernet.com/atlanta-weather-records or https://www.wunderground.com/history/airport/KFTY/2015/7/1/CustomHistory.html . You can use R if you’d like, but it’s straightforward enough that an Excel spreadsheet can easily do the job too

**Response –**

To find the initial critical value (C) and threshold(T) values, temperatures of July month are used for the year 1996. July 1996 temperatures have standard deviation of 4.9. Hence using the empirical formulae, C value of 2.5 and T value of 20 are chosen as initial values of T. Upon running multiple iterations with C and T values, it looks like C=5 and T =30 gives a better fit for the end of summer data as seen in the plot.

1. Use a CUSUM approach to make a judgment of whether Atlanta’s summer climate has gotten warmer in that time (and if so, when).

**Response –**

Change patterns in Atlanta’s weather are only visible if low values of threshold(T) are used. For a T value of 2, the change happens in the year 2000. For a T value of 5, the change happens in 2012.