HOMEWORK 3. ISYE 6501

Guillermo de la Hera Casado

September 8th, 2019

Question 5.1. Grubbs.test for outliers

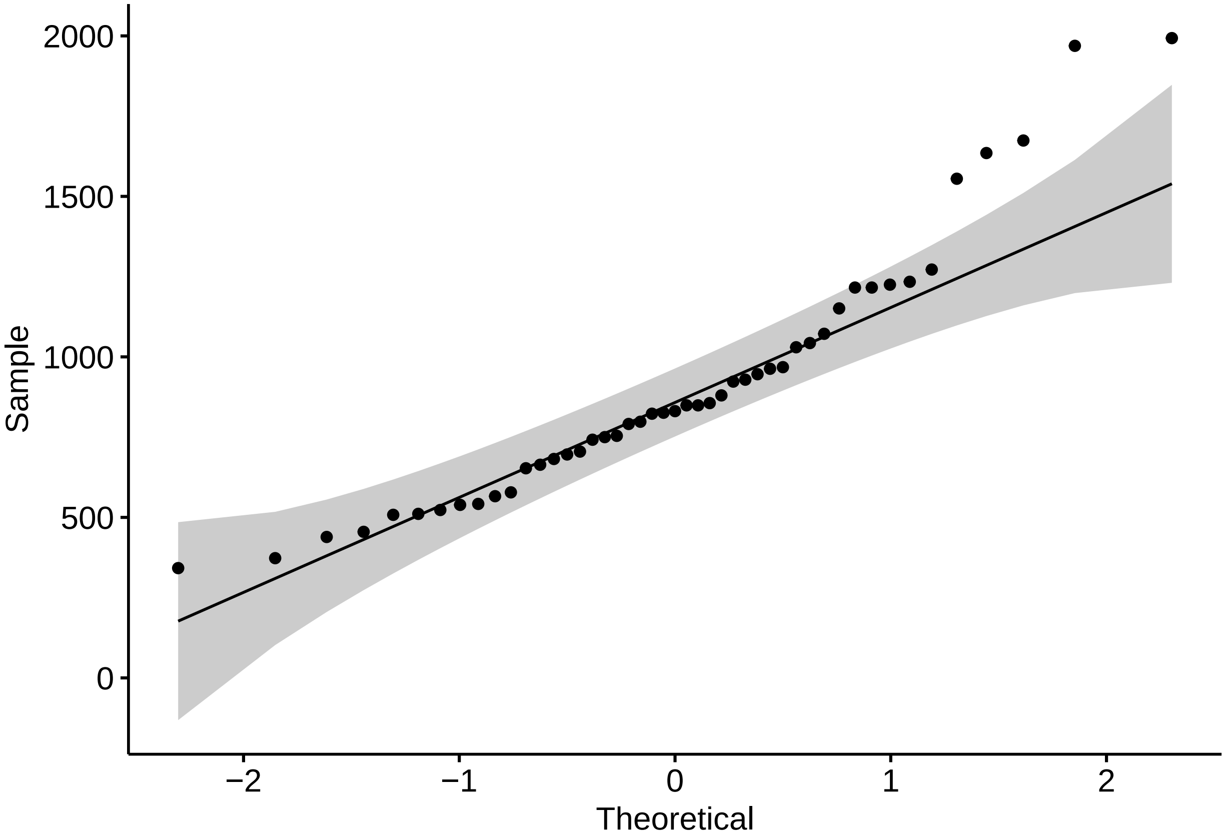
Grubbs' test is useful to identify outliers in normally distributed datasets, but can provide false results when tested on non-normal distributions.

Let's get the ball rolling:

```
usCrime = read.table("uscrime.txt", header = TRUE, sep = "\t")
```

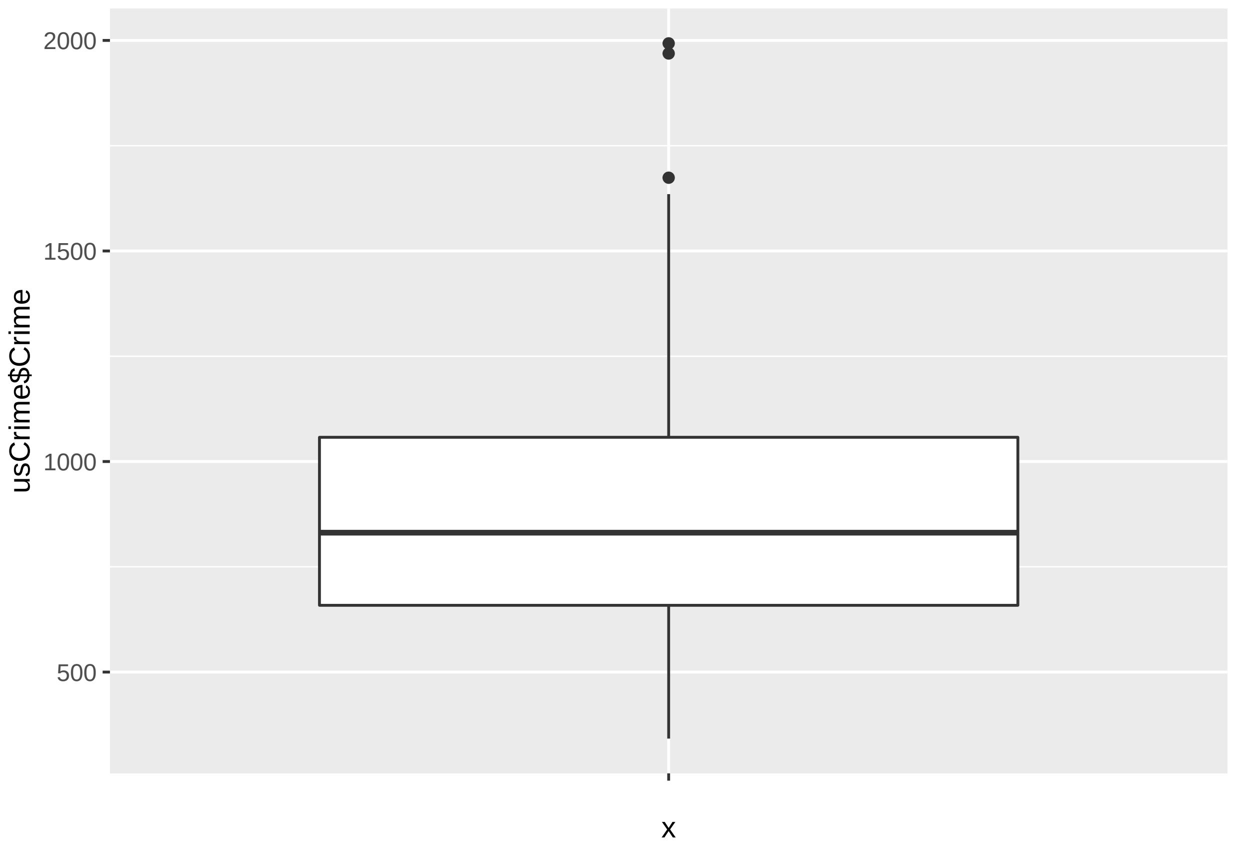
The first step will be then to identify visually whether our data is normally distributed and where the outliers are. For that, we will use first a Q-Q plot (quantile-quantile plot) that draws the correlation between a given sample and the normal distribution, with a 45-degree reference line:

```
ggqqplot(usCrime$Crime)
```



The data looks normally distributed but we can also to visualize few outliers. Let's check them even closely with a boxplot:

```
ggplot(data=usCrime, aes(x="", y = usCrime$Crime)) + geom_boxplot()
```



There is at least one data point scoring outside of the Whisker, and therefore being far away from *3rd quartile + 1.5 IQR* potentially being an outlier.

Let's formalize whether those points are formally outliers or not by using *Grubbs.test*. We will use *type=10* meaning that we want to validate whether the sample contans one outlier statistically different than the other values, in one tail:

- The null hypothesis is that value 1993 **is not** an outlier
- The alternative hypothesis is that value 1993 **is** an outlier

```
print(grubbs.test(usCrime$Crime, type = 10, two.sided = FALSE))
```

```
##
##  Grubbs test for one outlier
##
## data:  usCrime$Crime
## G = 2.81287, U = 0.82426, p-value = 0.07887
## alternative hypothesis: highest value 1993 is an outlier
```

The *p value* comes at **0.079**. There could be many possible scenarios here, depending on the desired cutoff, but let's mention two:

- If we wanted to know at 90% confidence level whether our observation is an outlier, then we would reject the null hypothesis, since $0.079 < 0.1$. At 90% confidence level, we would consider our point to be an outlier.
- If we wanted to use 95% confidence level, then we would not reject the null hypothesis, since 0.079 is not smaller than 0.05. At 95% confidence level, we wouldn't consider our point to be an outlier.

Question 6.1. Change detection model usefulness and application of CUSUM

I found an interesting article in Oxford Academic, where CUSUM charting is used to assess doctor's performance of consecutive procedures (pancreatography, renal and breast biopsies, etc).

- At acceptable levels of performance the CUSUM curve is flat.
- At unacceptable levels of performance, the curve slopes upward crossing the decision interval and providing early warning of adverse trend.

Doctors where actually happy with CUSUM as self-assessment tool, specially trainees who monitored their progress on learning new skills, flattening their CUSUM curves.

- S value would be chosen from analysis on historical data, understanding the fluctuation of records and seeing visually which S value triggers meaningul changes in performance.
- C value would then be adjusted in order to play with the sensitivity of the model, lower values make the model respond faster but also raises risk of false positives.

More info here, from Academic OUP

Question 6.2. CUSUM exercise

End of unofficial summer for Atlanta

Full dataset and visualizations in the spreadsheet, tab: *unofficial summer ends.* This exercise was resolved with the following steps:

- For each combination mm-dd, I made the average across the years.
- I plotted visually the averaged daily high per day, to identify where does the change start. Aug 30th breaks the range of the previous days, falling **up to 85.8**. From there the value of the observations only get lower.
- I computed Mu (average of the values before change, so before Aug 30th). **Mu = 88.8**
- I calculated the *S values* and plotted the control chart.
- I found that the combination **S = 6 and C = 0.5** makes Aug 30th reach the critical value as we want, but also avoids mid-July becoming a false positive (e.g. July 15th with temperature only at 87)

Did Atlanta's summmer get warmer from 1996 to 2015?

Full dataset and visualizations in the spreadsheet, tab *summer climate warmer.* Key actions as below:

- As we inferred from the previous exercise that unofficial summer ends on Aug 30th, I have sliced the data taking only observations from July 1st to Aug 30th) as an objective summer period.
- The second step was to average the columns, so that we get one avg value per year.
- After that I visualized the records. 0nly by 2011 we have a value that is higher than what is achieved in the previous local maximums (92.7). However from 2012 to 2014 we don't see particularly higher values, they fall within 1996-2010 range. **Therefore we cannot conclude that there is a change in behavior, neither that summers are getting warmer.**. The main consequence for us is that our **CUSUM model should not make 2011 cross the critical threshold.**
- I computed Mu (average of values before change, so 1996-2010).
- I calculated the *S values* and plotted the control chart.
- I used the combination **S = 0.6 and C = 1.2**, so that 2011 doesn't cross the critical value. High C values prevents from getting false positives and make the model less sensitive, that is exactly what we want.