Homework3

Pablo Diaz

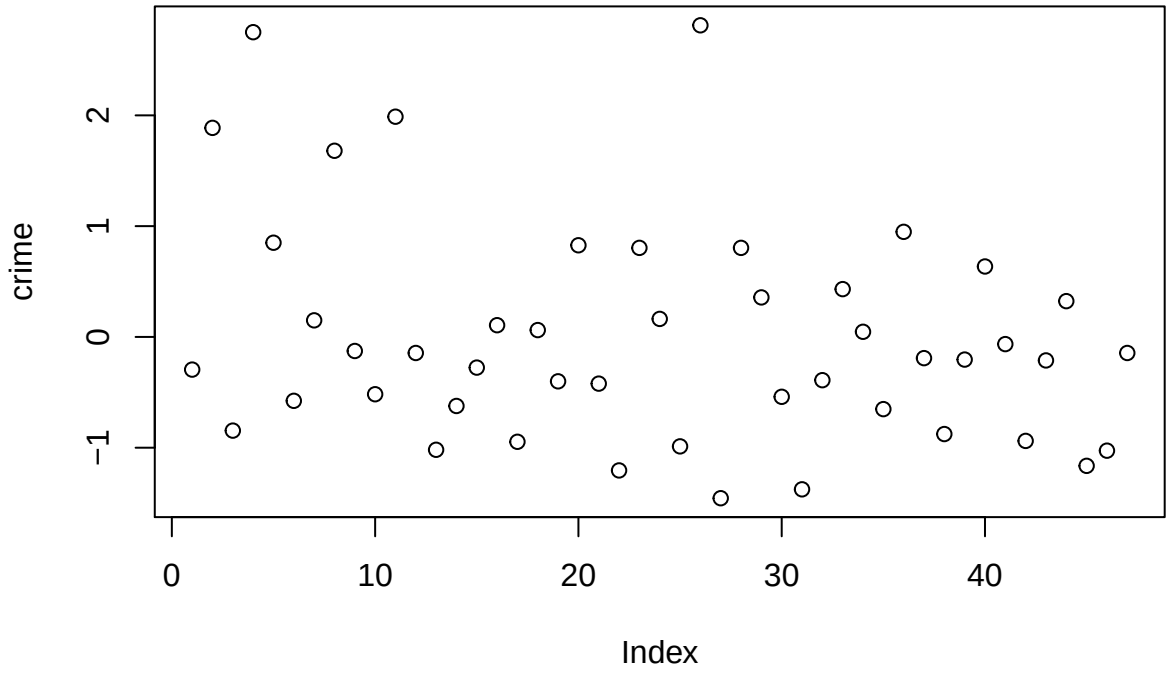
1/26/2019

Contents

Question 5.1

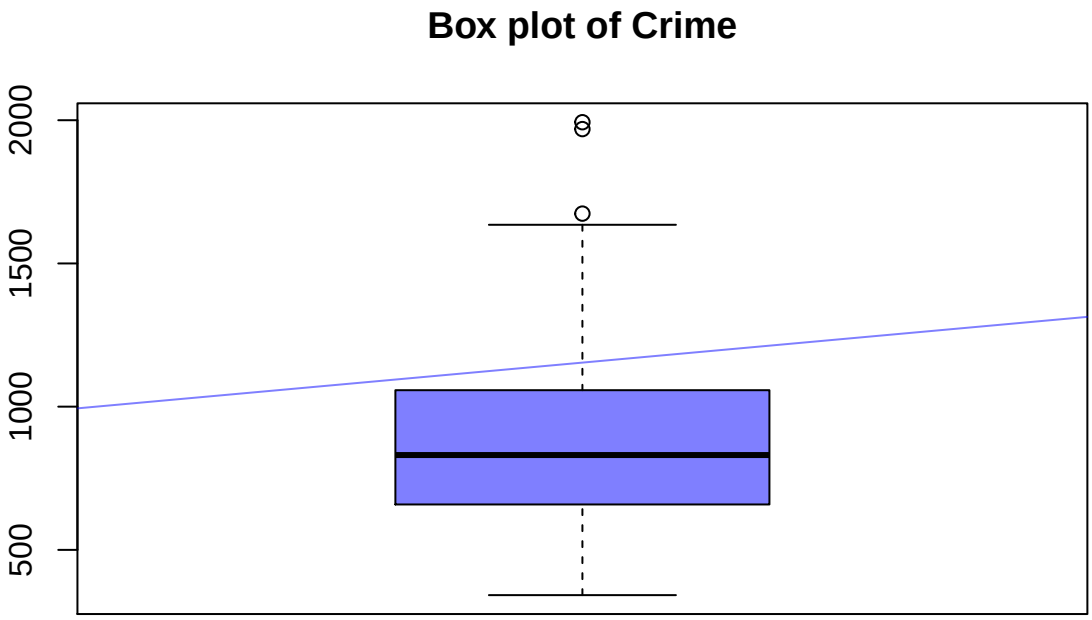
Use crime.data to test whether there are any outlier in the last column. Use the `grubbs.test` in the outliers package in R.

```
library(outliers)
set.seed(42) #set the seed
#import the data
uscrime <- read.delim("~/Documents/R/GeorgiaTech/DataPreparation/uscrime.txt")
crime <- scale(uscrime$Crime)
# before using the grubbs function it is a good idea to plot the points
plot(crime)
```



```r
# it is a simple plot, but I can observe that there are two points that are higher than most points
# I cannot see that there are some minium points that could be outliers.
#so let's test other plotting methods

# with this whiskers plot, I can confirm that there are 2 definite
#outliers and one other possibe outlier.
boxplot(uscrime$Crime,col=rgb(0,0,1,0.5), main="Box plot of Crime")
# This graph shows that most of the data behaves normal and shows 5 possible outliers.
qqline(uscrime$Crime,col=rgb(0,0,1,0.5))
```




```r
# now we can apply the grubbs test
outliers <- grubbs.test(uscrime$Crime, type = 10) # test for 1 possible outlier.
print (outliers)
```

```
##
##  Grubbs test for one outlier
##
## data:  uscrime$Crime
## G = 2.81290, U = 0.82426, p-value = 0.07887
## alternative hypothesis: highest value 1993 is an outlier
```

```r
# test the other side
outliers <- grubbs.test(uscrime$Crime, type = 10, opposite=TRUE) # test for 1 possible outlier.
# p-value = 1, no outlier.
print (outliers)
```

```
##
##  Grubbs test for one outlier
##
## data:  uscrime$Crime
## G = 1.45590, U = 0.95292, p-value = 1
## alternative hypothesis: lowest value 342 is an outlier
```

```r
# highest value 1993 is an outlier
# remove first outlier
data <- uscrime$Crime[-match(c(1993), uscrime$Crime)]
```

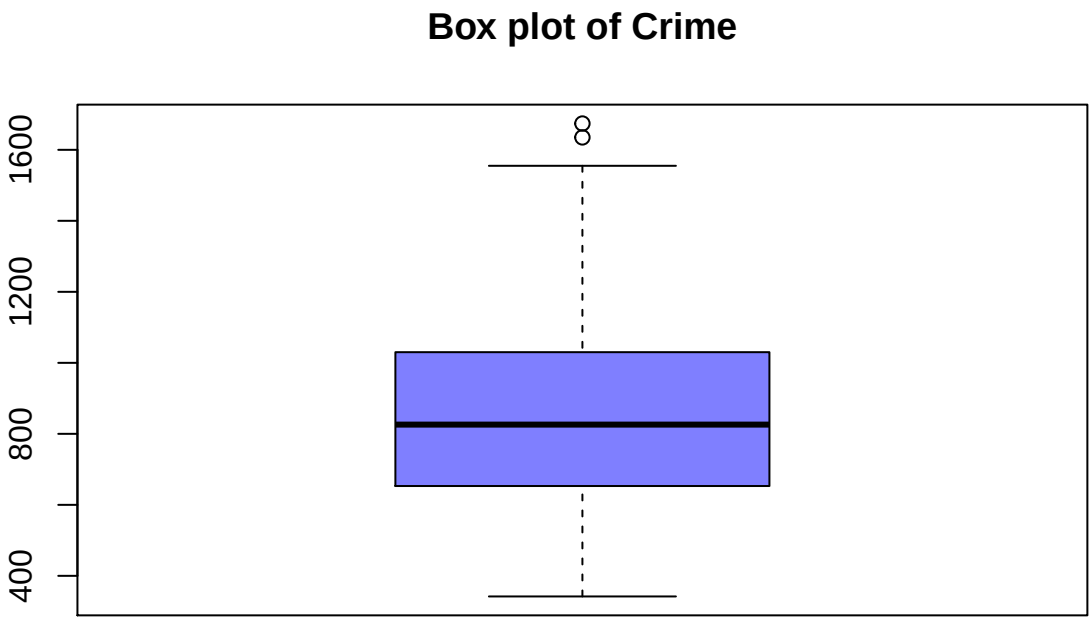
```
# test again
outliers <- grubbs.test(data, type = 10)
# p-value is not high enough, highest value 1969 is an outlier
print (outliers)
```

```
##
##  Grubbs test for one outlier
##
## data:  data
## G = 3.06340, U = 0.78682, p-value = 0.02848
## alternative hypothesis: highest value 1969 is an outlier
```

```
data <- data[-match(c(1969), data)]
# test again, p-value here is high enough
outliers <- grubbs.test(data, type = 10)
print (outliers)
```

```
##
##  Grubbs test for one outlier
##
## data:  data
## G = 2.56460, U = 0.84712, p-value = 0.1781
## alternative hypothesis: highest value 1674 is an outlier
```

```
# now show the box and whiskers plot without outliers
boxplot(data,col=rgb(0,0,1,0.5), main="Box plot of Crime")
```




```
# what else could be done? Test other outlier methods and see the results.
library(EnvStats)
```

```
##
## Attaching package: 'EnvStats'
```

```
## The following objects are masked from 'package:stats':
##
##     predict, predict.lm
```

```
## The following object is masked from 'package:base':
##
##     print.default
```

```
# here k is the potential number of outliers
rTest_1 <- rosnerTest(uscrime$Crime, k = 2, alpha = 0.05, warn = TRUE)
# no outliers found!
print (rTest_1)
```

```
##
## Results of Outlier Test
## -------------------------
##
## Test Method:                     Rosner's Test for Outliers
##
## Hypothesized Distribution:       Normal
##
## Data:                            uscrime$Crime
##
## Sample Size:                     47
##
## Test Statistics:                 R.1 = 2.812874
##                                  R.2 = 3.063425
##
## Test Statistic Parameter:        k = 2
##
## Alternative Hypothesis:          Up to 2 observations are not
##                                  from the same Distribution.
##
## Type I Error:                    5%
##
## Number of Outliers Detected:     0
##
##   i   Mean.i     SD.i Value Obs.Num    R.i+1 lambda.i+1 Outlier
## 1 0 905.0851 386.7627  1993      26 2.812874   3.103243   FALSE
## 2 1 881.4348 355.0161  1969       4 3.063425   3.094456   FALSE
```

```
rTest_2 <- rosnerTest(uscrime$Crime, k = 3, alpha = 0.05, warn = TRUE)
# no outliers found!
print (rTest_2)
```

```
##
## Results of Outlier Test
## -------------------------
##
## Test Method:                     Rosner's Test for Outliers
##
## Hypothesized Distribution:       Normal
##
## Data:                            uscrime$Crime
##
## Sample Size:                     47
##
## Test Statistics:                 R.1 = 2.812874
##                                  R.2 = 3.063425
```

```
##                                  R.3 = 2.564571
##
## Test Statistic Parameter:        k = 3
##
## Alternative Hypothesis:          Up to 3 observations are not
##                                  from the same Distribution.
##
## Type I Error:                    5%
##
## Number of Outliers Detected:     0
##
##   i   Mean.i     SD.i Value Obs.Num    R.i+1 lambda.i+1 Outlier
## 1 0 905.0851 386.7627  1993      26 2.812874   3.103243   FALSE
## 2 1 881.4348 355.0161  1969       4 3.063425   3.094456   FALSE
## 3 2 857.2667 318.4678  1674      11 2.564571   3.085425   FALSE
```

Analysis

First I made some plots to see how the data was behaving. I saw at least three possible outliers in the highest parts of the data. After removing 2 of them, I plotted the whiskers box and saw that it showed one possible outlier. Since the data is only 47 points, it is probably that the graph and the grubbs test will always say that there is an outlier. So I just removed two of them according to the `p-values`. Also before applying grubbs test I saw if the data had some nomal behavior through the qqplot.

I didn't used a normality test because I didn't found one that was relaible. Normality tests don't do what most think they do. Shapiro's test, Anderson Darling, and others are null hypothesis tests AGAINST the the assumption of normality. These should not be used to determine whether to use normal theory statistical procedures. In fact they are of virtually no value to the data analyst. Under what conditions are we interested in rejecting the null hypothesis that the data are normally distributed?

To test another method to detect outliers, I chose to perform Rosner's generalized extreme Studentized deviate test for up to k potential outliers in a dataset, assuming the data without any outliers come from a normal (Gaussian) distribution. This test said that there were no outliers, so it was not very helpful. Notice that I tested it with the original data and not with the data without outliers.

Question 6.1

Sea level decrease/increase is relevant to measure global warming and relevant to measure possible natural disasters as high tides or tsunamis (when the water is pulled). CUSUM can be used to measure and monitor when abrupt changes occur in the sea level in certain threshold, possibly showing evidence of a natural disaster is coming up. Sea level varies naturally, so in this case the threshold should be high enough to only detect abrupt changes in the sea level, a choice could be T = 10 (standard deviations) and C = half standard deviation. This is just an aproximation, but in real world examples there is more information to approximate better this parameters.

Question 6.2

1. identify when unofficial summer ends (cool off)
2. judgment of whether Atlanta's summer climate has gotten warmer in that time (and if so, when).

Use the average for each day and do CUSUM on that Do CUSUM for each year, then find average change detection date.

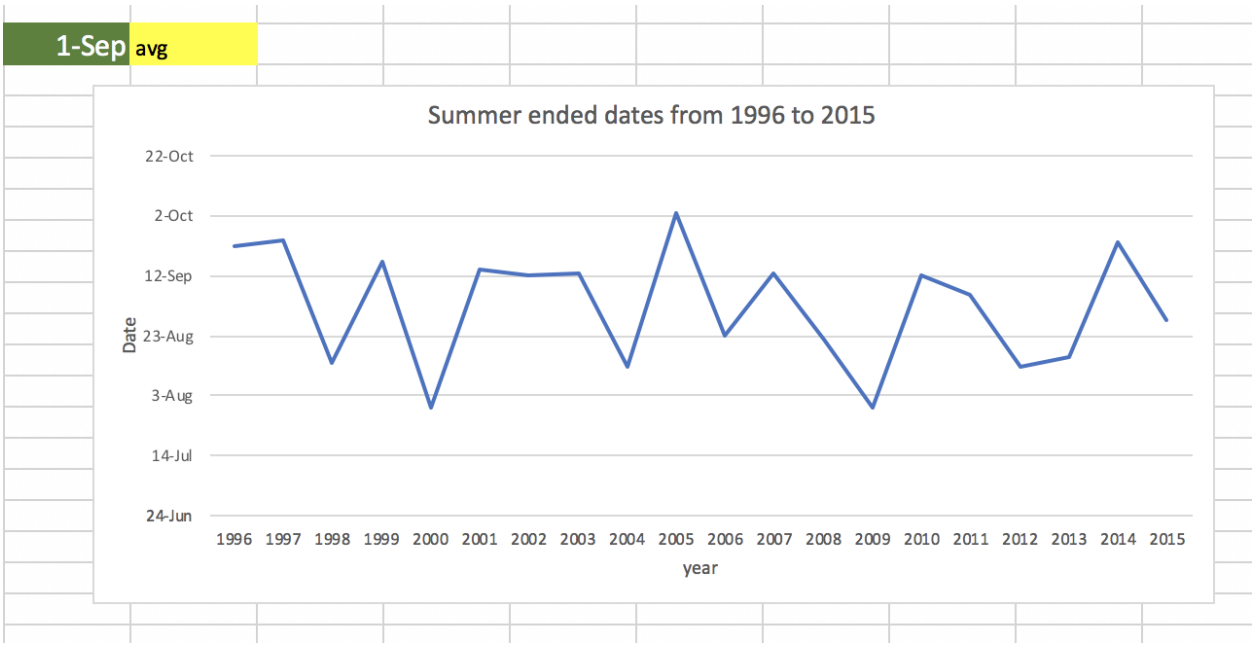


Figure 1: Summer ending dates C=5, T=50

Analysis 6.2.1

First the equation to detect decrease is the following

x_t = is the observed value at time t.

μ = the expected value of the observations x.

x_t - μ = how much above/below expected the observation is at time t.

$S_t = \max\{0, S_t+(\mu\text{-}x_t\text{-C})\}$ (for decrease)

The change is detected when S_t>=T. C and T are model paramters.

First I calculated the μ from July to mid August because we know for sure that summer doesn't end in before mid August. (mu should be the average when you are sure no change had occured). Then applied CUSUM to each day in each year, making a matrix in the spreadsheet. Then I made a function to find the index when the T value was reached, I mean when the algorithm detected a decrease according to the parameters.

I tested multiple parameters to find out that the average date is very sensible to the T value used as the paramter model. The C value was not very senitive, and by sensitive I mean that a small change in these parameters could make a big change in the average date when summer ended. With a high threshold of T the we detect changes slower but is less likely to falsely detect changes. In this example, that is our goal because it is more important to have a more accurate date. For the critical value I left the default option of 5.

For example I tested C = 5 and T = 50 and the result was that summer ends on 1 Sept (Figure 1). When I tested C = 100 and T = 50, the date only moved 2 days (Figure 2) and when I tested C = 5 and T = 100, the date moved 16 days, so the CUSUM is more sensitive to the T parameter and there is no single right result.

The final result is that unofficial summer ends around 1-Sep according to the data we have from 1996 to 2015.

Analysis 6.2.2

For this problem we want to solve the question if summer is longer than before

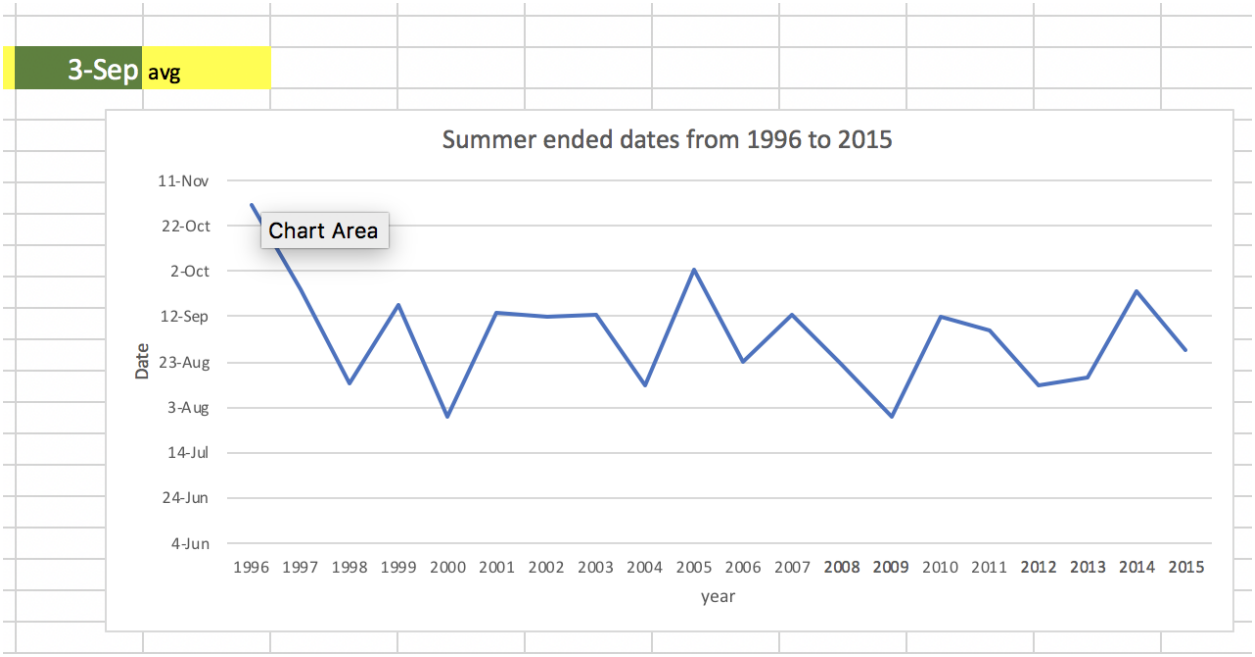


Figure 2: Summer ending dates C=100, T=50

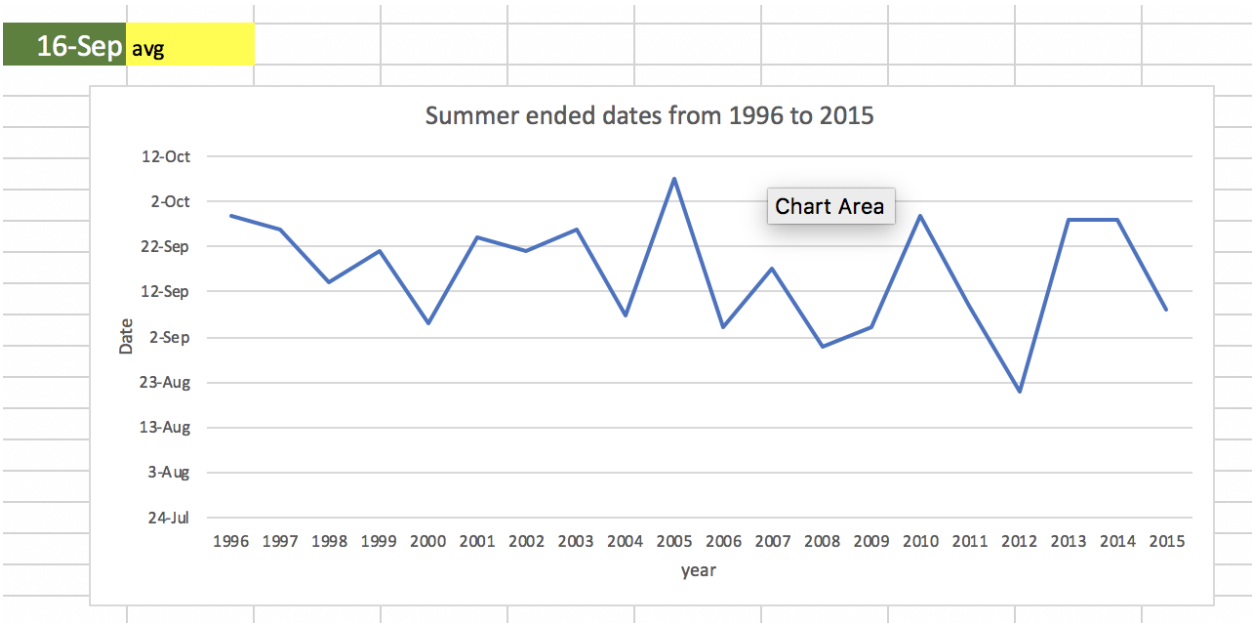


Figure 3: Summer ending dates C=5, T=100

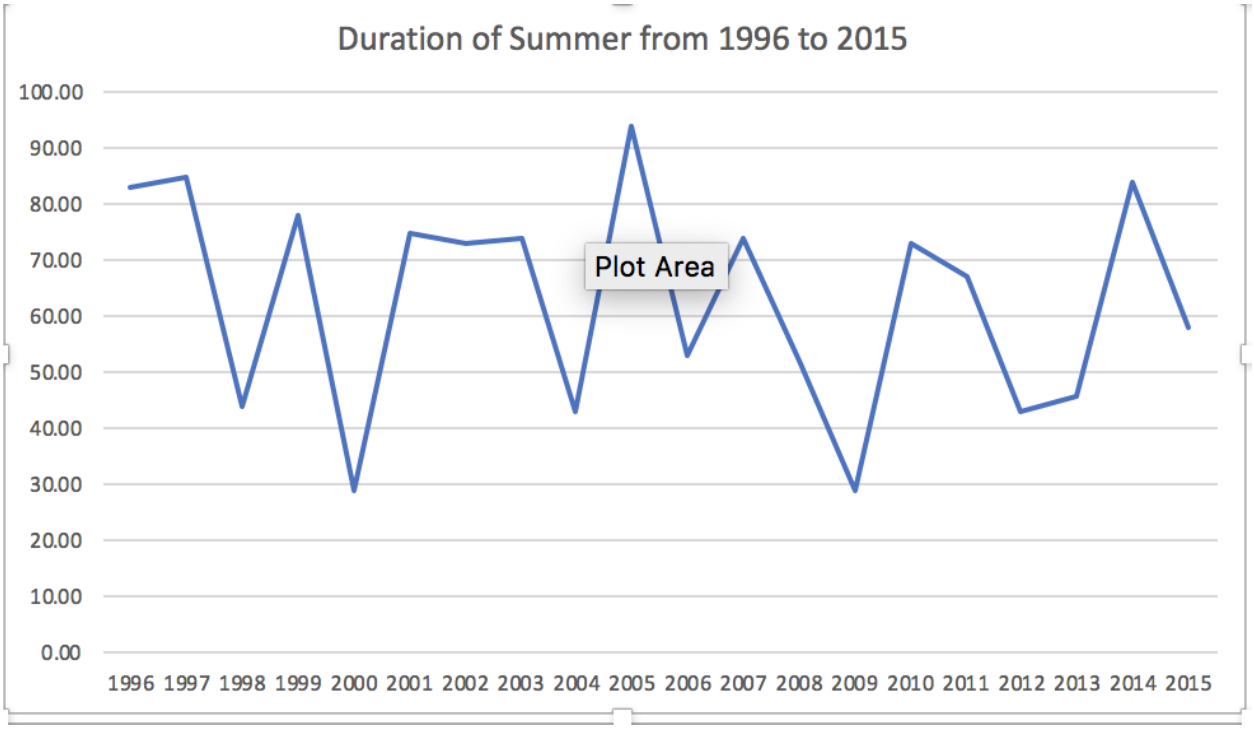


Figure 4: Summer duration by year

So basically we have to detect a change but now it will be what year summer lasted longer, not in a day/month.

The same equation is used to apply CUSUM but now it will be according the duration of summer each year.

In figure 4 it is shown the duration of summer across the years, as the graph shows there seems not to be a pattern, it changes a lot from year to year. There must be other factors that affect the duration of the summer each year.

In Figure 5, a made a small table showing how the paramters could change the year that summer changed significantly. The hypothesis that summer climate has gotten warmer in that time cannot be confirmed or denied with my results. The result is too sensitive but 2005 seems to be the year when summer lasted longer but then in 2009 was the year that lasted the least days.

C	T	Year changed
5	25	2005
5	10	1999
5	12	2012
10	12	2005
10	15	Not changed
10	25	Not changed

Figure 5: Summer duration paramters