ISyE6501_HW3

JinyuHuang

9/8/2019

5.1

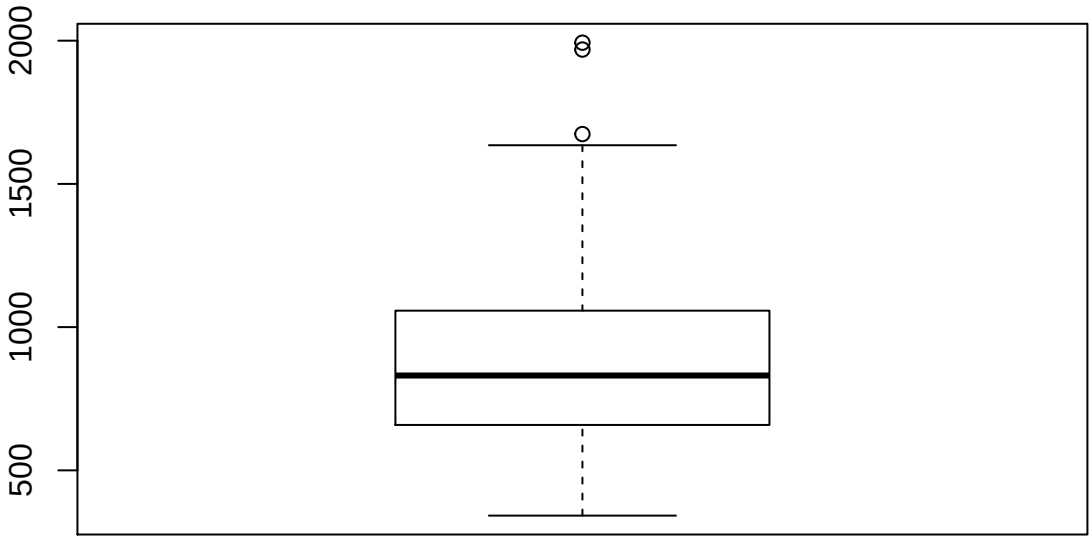
```
uscrime <- read.table("/Users/luckyfisher/Desktop/Courses/ISyE\ 6501/hw3/uscrime.txt",
                      header = TRUE)
str(uscrime)
```

```
## 'data.frame':    47 obs. of  16 variables:
##  $ M     : num  15.1 14.3 14.2 13.6 14.1 12.1 12.7 13.1 15.7 14 ...
##  $ So    : int  1 0 1 0 0 0 1 1 1 0 ...
##  $ Ed    : num  9.1 11.3 8.9 12.1 12.1 11 11.1 10.9 9 11.8 ...
##  $ Po1   : num  5.8 10.3 4.5 14.9 10.9 11.8 8.2 11.5 6.5 7.1 ...
##  $ Po2   : num  5.6 9.5 4.4 14.1 10.1 11.5 7.9 10.9 6.2 6.8 ...
##  $ LF    : num  0.51 0.583 0.533 0.577 0.591 0.547 0.519 0.542 0.553 0.632 ...
##  $ M.F   : num  95 101.2 96.9 99.4 98.5 ...
##  $ Pop   : int  33 13 18 157 18 25 4 50 39 7 ...
##  $ NW    : num  30.1 10.2 21.9 8 3 4.4 13.9 17.9 28.6 1.5 ...
##  $ U1    : num  0.108 0.096 0.094 0.102 0.091 0.084 0.097 0.079 0.081 0.1 ...
##  $ U2    : num  4.1 3.6 3.3 3.9 2 2.9 3.8 3.5 2.8 2.4 ...
##  $ Wealth: int  3940 5570 3180 6730 5780 6890 6200 4720 4210 5260 ...
##  $ Ineq  : num  26.1 19.4 25 16.7 17.4 12.6 16.8 20.6 23.9 17.4 ...
##  $ Prob  : num  0.0846 0.0296 0.0834 0.0158 0.0414 ...
##  $ Time  : num  26.2 25.3 24.3 29.9 21.3 ...
##  $ Crime : int  791 1635 578 1969 1234 682 963 1555 856 705 ...
```

```
library(ggplot2)
crime <- uscrime$Crime
str(crime)
```

```
##  int [1:47] 791 1635 578 1969 1234 682 963 1555 856 705 ...
```

```
boxplot(crime)
```



Outlier detection of the upper side

We apply Grubbs Test to check whether there is suspicious outlier in the vector of "Crime". For the test of largest outlier, the Null Hypothesis is that "highest value 1993 is not an outlier". The G value we retrieve from the model is 2.812874, which is lower than G95(45), approximately equal to 2.914. Based on the statistic, we tend to believe that outlier doesn't exist in the "Crime" data. When we turn to the P-value: 0.07887486 in this case, it's higher than 0.05, which means by a 95% confidence interval, the Null Hypothesis holds.

```
library(outliers)
set.seed(1234)
x = crime
detect_upper <- grubbs.test(crime, type = 10, opposite = FALSE)
detect_upper$statistic
```

```
##        G        U
## 2.812874 0.824255
```

```
detect_upper$p.value
```

```
## [1] 0.07887486
```

```
detect_upper$alternative
```

```
## [1] "highest value 1993 is an outlier"
```

Outlier detection of the lower side

We also detect whether the smallest value is an outlier. The P-value 1 indicates that the Null Hypothesis "lowest value 342 is an outlier" will hold.

```
library(outliers)
set.seed(1234)
x = crime
detect_lower <- grubbs.test(crime, type = 10, opposite = TRUE)
detect_lower$statistic
```

```
##         G         U
## 1.4558930 0.9529195
```

```
detect_lower$p.value
```

```
## [1] 1
```

```
detect_lower$alternative
```

```
## [1] "lowest value 342 is an outlier"
```

6.1

Assume transportation management office would like to test the influence that dentisy of fog has on the responding ability of drivers. By identifying the marginal change of average responding time of drivers by unit increase in the density of fog, the department hopes to set a more affective speed limit in order to prevent traffic accidents in foggy environment. In this case, the CUSUM methodology can help us identify the significant changes that occur during the dynamic procedure by taking in numeric parameters that indicate density of fog and drivers' responding time as dependent variable, which may be retrived from lab experiments,.

The determination of critical value (C) and threshold (T) needs to go through trial and error. We also need to look into the realistic influence of change in drivers' responding time on the risk of traffic accidents. At the meantime, we may want the CUSUM model to be more sensitive in this case as the cost of sudden increase in responding time is people's lives. That's why we will lower C and lower T to make the CUSUM more predictive.

6.2

```
temps <- read.table("/Users/luckyfisher/Desktop/Courses/ISyE\ 6501/hw3/temps.txt",
                    header = TRUE)
str(temps)
```

```
## 'data.frame':    123 obs. of  21 variables:
##  $ DAY  : Factor w/ 123 levels "1-Aug","1-Jul",..: 2 46 90 101 105 109 113 117 121 6 ...
##  $ X1996: int  98 97 97 90 89 93 93 91 93 93 ...
##  $ X1997: int  86 90 93 91 84 84 75 87 84 87 ...
##  $ X1998: int  91 88 91 91 91 89 93 95 95 91 ...
##  $ X1999: int  84 82 87 88 90 91 82 86 87 87 ...
##  $ X2000: int  89 91 93 95 96 96 96 91 96 99 ...
##  $ X2001: int  84 87 87 84 86 87 87 89 91 87 ...
##  $ X2002: int  90 90 87 89 93 93 89 89 90 91 ...
##  $ X2003: int  73 81 87 86 80 84 87 90 89 84 ...
##  $ X2004: int  82 81 86 88 90 90 89 87 88 89 ...
##  $ X2005: int  91 89 86 86 89 82 76 88 89 78 ...
##  $ X2006: int  93 93 93 91 90 81 80 82 84 84 ...
##  $ X2007: int  95 85 82 86 88 87 82 82 89 86 ...
##  $ X2008: int  85 87 91 90 88 82 88 90 89 87 ...
##  $ X2009: int  95 90 89 91 80 87 86 82 84 84 ...
##  $ X2010: int  87 84 83 85 88 89 94 97 96 90 ...
##  $ X2011: int  92 94 95 92 90 90 94 94 91 92 ...
##  $ X2012: int  105 93 99 98 100 98 93 95 97 95 ...
##  $ X2013: int  82 85 76 77 83 83 79 88 88 87 ...
##  $ X2014: int  90 93 87 84 86 87 89 90 90 87 ...
##  $ X2015: int  85 87 79 85 84 84 90 90 91 93 ...
```

We manipulate Excel to build up CUSUM and detect cooling change in temperature. C is given value 4, T is given value 25. According to observation, the significant cooling point happened in each year within the date range of 6th Aug. to 4th Oct., which is in general accords with common sense of summer end. 2 out the 20 years had the cooling point in August, 2 out of the 20 years are detected cooling point in October, the rest 16 years have their estimated cooling point distributed in September(Details please refer to file: "temps_analysis").

Furthermore, we retrieve the data of cooling dates and their corresponding temperatures through 1996 to 2015. By applying CUSUM to the data(C=1,T=5), we can tell that year 2001 is a change point which indicates significant increase in the temperature at the summer end.