Outlier Detection

Kent Ng

May 26, 2018

Let's use Grubbs' test to determine whether there are any outliers in the last column of the crime data from file uscrime.txt (http://www.statsci.org/data/general/uscrime.txt (http://www.statsci.org/data/general/uscrime.txt), description at http://www.statsci.org/data/general/uscrime.html (http://www.statsci.org/data/general/uscrime.html)).

Let's load the "outliers" library. As best practice, let's also set seed value.

```
library("outliers")
set.seed(42)
```

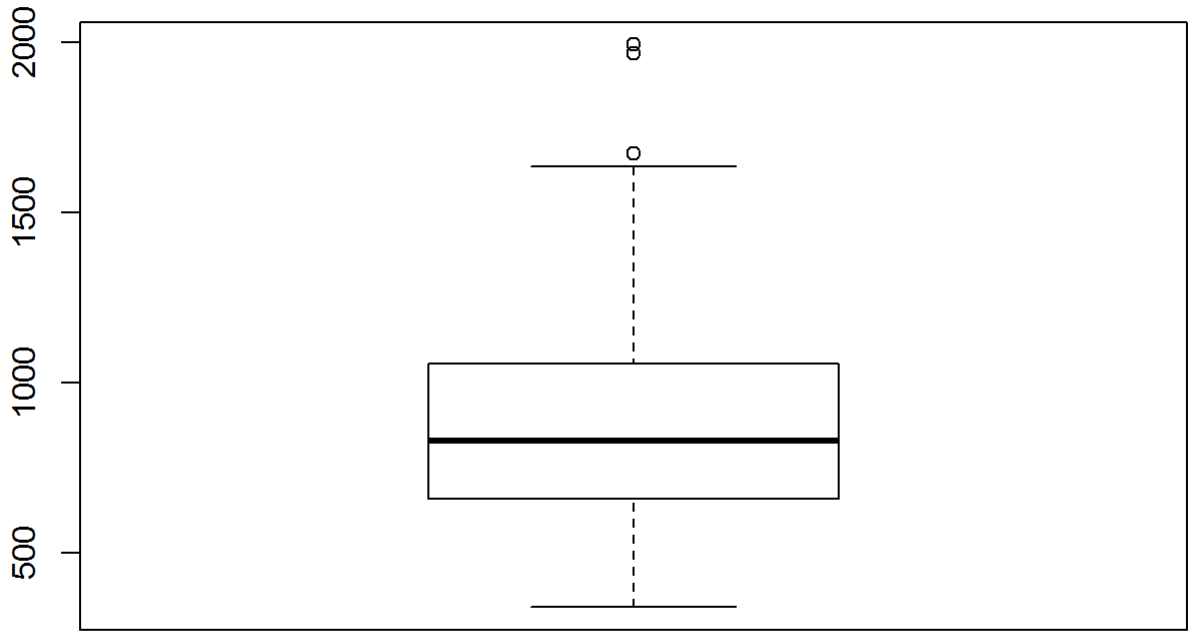
Next, let's read the values from "5.1uscrimeSummer2018.txt" into a dataframe data_5

```
data_5<-read.table("5.1uscrimeSummer2018.txt",header=TRUE)
```

Before we use the grubbs test to test for outliers, let's plot and visualize our data.

```
hist(data_5[,16])
```

```
boxplot(data_5[,16])
```



Now, as requested by the question, let's use the grubbs.test function to determine whether there are any outliers in the last column of data_5. There are different types of grubbs tests. Based on the histogram and boxplot above, we can see that our potential outliers are all on the upper end. Therefore, we know that a one-tailed test would suffice in this situation.

Let's first use type=10 (default) to determine if the data contains one outlier that is statistically different than the other values. The null hypothesis is that there is no outliers in the data.

```
v16<-data_5[,16]
grubbs.test(v16, two.sided=F, type=10)
```

```
##
##  Grubbs test for one outlier
##
## data:  v16
## G = 2.81290, U = 0.82426, p-value = 0.07887
## alternative hypothesis: highest value 1993 is an outlier
```

This Grubbs' test tested the alternative hypothesis that 1993 is an outlier. However, the p value for the test is 0.079 and the typical alpha value is 0.05 (cutoff for significance). Thus, we fail to reject the null hypothesis and can conclude that there is no outlier in our dataset. Note that the 0.05 cutoff is arbitrary and so technically one could still argue that 1993 could potentially be an outlier.

Let’s try other types of Grubbs test

```
# check the opposite tail for any outliers - Result: Failed to reject null hypothesis
grubbs.test(v16, two.sided=F, type=10, opposite=TRUE)
```

```
##
##  Grubbs test for one outlier
##
## data:  v16
## G = 1.45590, U = 0.95292, p-value = 1
## alternative hypothesis: lowest value 342 is an outlier
```

```
# Use Type 11 to check if the lowest and highest values are two outliers on opposite tails - Res
ult: Failed to reject null hypothesis
grubbs.test(v16, two.sided=F, type = 11)
```

```
##
##  Grubbs test for two opposite outliers
##
## data:  v16
## G = 4.26880, U = 0.78103, p-value = 1
## alternative hypothesis: 342 and 1993 are outliers
```

```
# Let's test for outliers using a two-tailed approach - Result: Failed to reject null hypothesis
grubbs.test(v16, two.sided=T, type=10)
```

```
##
##  Grubbs test for one outlier
##
## data:  v16
## G = 2.81290, U = 0.82426, p-value = 0.1577
## alternative hypothesis: highest value 1993 is an outlier
```

Based on our test results, we can conclude that there are no outliers in the last column of our data.