Crime_Outliers

Hugh McMurrain

—– HW2 5.1 —–

Clear Environment, load necessary libraries, load data and ensure it was read correctly.

```
rm(list=ls())

library(ggplot2)
library(outliers)

data <- read.table("C:/Users/tmcmu/Desktop/edX Courses/ISYE 6501/Homework
2/data 5.1/uscrime.txt", stringsAsFactors = FALSE, header = TRUE)

head(data)
```

```
##      M So   Ed  Po1  Po2    LF   M.F Pop   NW    U1  U2 Wealth Ineq
Prob
## 1 15.1  1  9.1  5.8  5.6 0.510  95.0  33 30.1 0.108 4.1   3940 26.1
0.084602
## 2 14.3  0 11.3 10.3  9.5 0.583 101.2  13 10.2 0.096 3.6   5570 19.4
0.029599
## 3 14.2  1  8.9  4.5  4.4 0.533  96.9  18 21.9 0.094 3.3   3180 25.0
0.083401
## 4 13.6  0 12.1 14.9 14.1 0.577  99.4 157  8.0 0.102 3.9   6730 16.7
0.015801
## 5 14.1  0 12.1 10.9 10.1 0.591  98.5  18  3.0 0.091 2.0   5780 17.4
0.041399
## 6 12.1  0 11.0 11.8 11.5 0.547  96.4  25  4.4 0.084 2.9   6890 12.6
0.034201
##      Time Crime
## 1 26.2011   791
## 2 25.2999  1635
## 3 24.3006   578
## 4 29.9012  1969
## 5 21.2998  1234
## 6 20.9995   682
```

“Crime” data - V16 - is the only column of interest

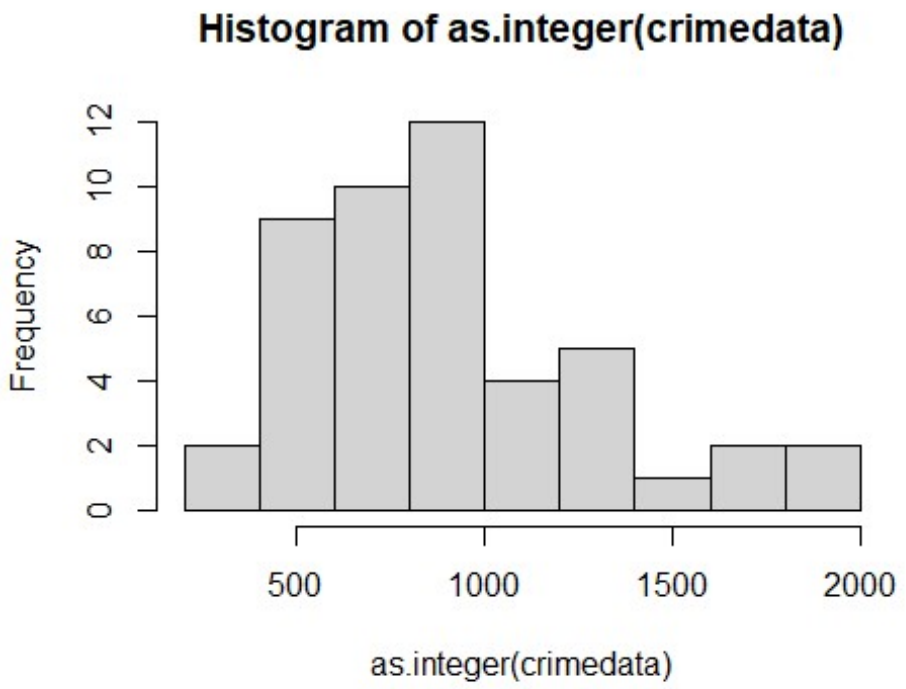
```
crimedata <- data[,"Crime"]
```

set seed

```
set.seed(22)
```

Check to see if Crime (V16) is normally distributed

```
hist(as.integer(crimedata), breaks = 6)
```



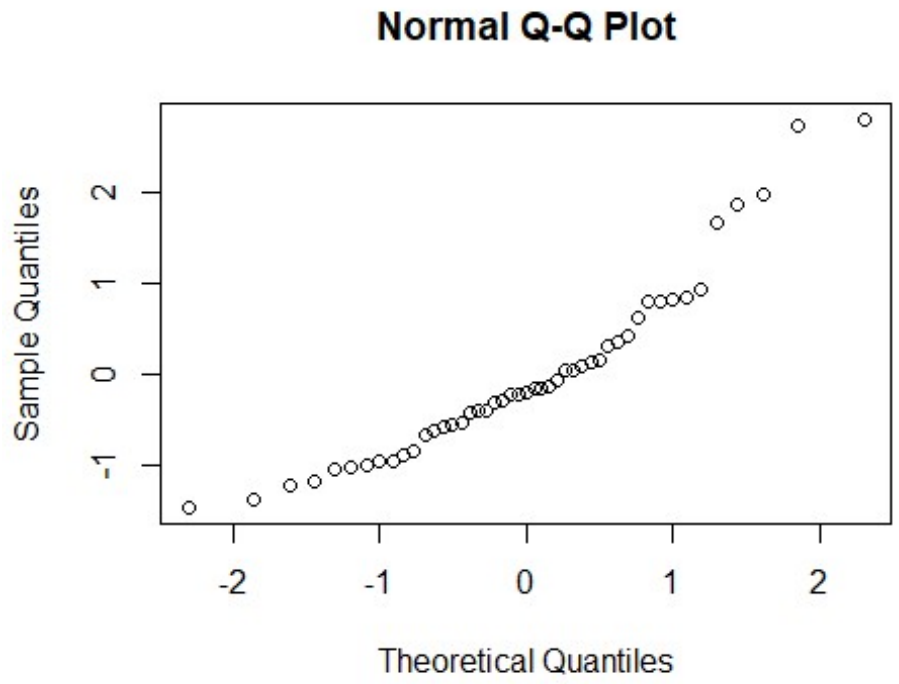

```
shapiro.test(crimedata)

##
##  Shapiro-Wilk normality test
##
## data:  crimedata
## W = 0.91273, p-value = 0.001882
```

Both the histogram and the shapiro test p-value (0.001882) show that the data is NOT normally distributed.

Let's use a qqnorm plot with SCALED data to determine if the middle of the data set is normally distributed.

```
qqnorm(scale(crimedata))
```



the QQnorm plot suggests that the "middle" of the data is approximately normally distributed, thus we can run a Grubbs' test. A Grubbs' test of type=11 will run the test for two outliers on opposite tails of the distribution.

```
CrimeModel <- grubbs.test(crimedata, type = 11)
CrimeModel

##
##  Grubbs test for two opposite outliers
##
## data:  crimedata
## G = 4.26877, U = 0.78103, p-value = 1
## alternative hypothesis: 342 and 1993 are outliers
```

P-value of 1 shows that one of the extremes, 342 or 1993, is an NOT an outlier.

Retry the grubbs test with a type of 10

```
CrimeModel <- grubbs.test(crimedata, type = 10)
CrimeModel

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

With a p-value of 0.07887, the 1993 data point could either be considered an outlier or not, depending on which threshold p-value we utilized for an outlier cutoff (0.05 vs. 0.10). Lets assume this data point is an outlier.

Now we will create a second dataset without the outlier to run a grubbs test on.

```
which.max(crimedata)

## [1] 26

CrimeModel_NoMax <- crimedata[-which.max(crimedata)]
grubbs.test(CrimeModel_NoMax, type = 10)

##
##  Grubbs test for one outlier
##
## data:  CrimeModel_NoMax
## G = 3.06343, U = 0.78682, p-value = 0.02848
## alternative hypothesis: highest value 1969 is an outlier
```

The low p-value < 0.05, which suggests that this city's crime is also an outlier. So let's remove it, and run the Grubb's test again.

```
which.max(CrimeModel_NoMax)

## [1] 4

CrimeModel3 <- CrimeModel_NoMax[-which.max(CrimeModel_NoMax)]
grubbs.test(CrimeModel3, type = 10)

##
##  Grubbs test for one outlier
##
## data:  CrimeModel3
## G = 2.56457, U = 0.84712, p-value = 0.1781
## alternative hypothesis: highest value 1674 is an outlier
```

The p-value (0.1781) is high enough that it isn't clear if the data point is an outlier. So let's stop there, having removed the top 2 highest potential outliers.

Now we can check the minimum data point, and since we've highlighted and removed the two maximum outliers, only the minimum data point is left to be tested. To do this we can use the OPPOSITE=TRUE parameter.

```
Low_OutlierTest <- grubbs.test(CrimeModel3, type=10, opposite = TRUE)
Low_OutlierTest

##
##  Grubbs test for one outlier
##
## data:  CrimeModel3
```

```
## G = 1.61796, U = 0.93915, p-value = 1
## alternative hypothesis: lowest value 342 is an outlier
```

With a p-value of 1, we can reasonably conclude that the minimum data point is NOT an outlier.