Boxplots

JuJu

## Factor Variables

R calls categorical or qualitative variables, factor variables. When reading CSV files, R makes character variables into factor variables by default. However, when a factor variable has been coded with numbers, R assumes that the variable is quantitative. The ***HTWT*** data shows how this can be a problem.

 htwt = read.csv("http://facweb1.redlands.edu/fac/jim\_bentley/Data/FYS04/HtWt.csv")
 summary(htwt)

## Height Weight Group
## Min. :51.0 Min. : 82.0 Min. :1.00
## 1st Qu.:56.0 1st Qu.:108.2 1st Qu.:1.00
## Median :59.5 Median :123.5 Median :2.00
## Mean :62.1 Mean :139.6 Mean :1.55
## 3rd Qu.:68.0 3rd Qu.:166.8 3rd Qu.:2.00
## Max. :79.0 Max. :228.0 Max. :2.00

Note that the variable ***Group*** has been treated as numeric. It turns out that this variable actually represents the sex of the individual and that males were coded as 1 and females as 2. We convert the numeric variable to a factor variable.

 is.numeric(htwt$Group)

## [1] TRUE

 is.factor(htwt$Group)

## [1] FALSE

 table(htwt$Group)

##
## 1 2
## 9 11

 htwt$Group = factor(htwt$Group, labels=c("Male","Female"))
 is.numeric(htwt$Group)

## [1] FALSE

 is.factor(htwt$Group)

## [1] TRUE

 summary(htwt$Group)

## Male Female
## 9 11

 table(htwt$Group)

##
## Male Female
## 9 11

## Simple Boxplots

Simple boxplots graphically represent the five number summary. The “Five Number Summary” handout shows how to compute these values. For the ***htwt*** data we see that the values are 82, 106.5, 123.5, 174.5, and 228.

R makes the generation of boxplots simple.

 # Base R
 boxplot(htwt$Weight)
 # Lattice
 p\_load(lattice)



 bwplot(~Weight, data=htwt)



 # ggplot2
 p\_load(ggplot2)
 ggplot(htwt, aes(' ',Weight)) + geom\_boxplot() + xlab("")



The top of the whisker is the maximum. The top of the box is the upper quartile. The median is the line (or black dot) in the middle of the box. The bottom of the box is the lower quartile. Finally, the bottom of the whisker is the minimum.

## Possible Outliers

Possible outliers may be found by their position relative to the upper and lower quartiles. We define the lower adjacent value by the smallest observed value greater than **Q1 - 1.5*IQR and the upper adjacent value by the largest observed value less than Q3 + 1.5*IQR**. These values become the new tips of the whiskers. Observed values outside of the fences are considered to be possible outliers.

For the ***htwt*** data we can compute the adjacent values and check for outliers.

 q3 = 174.5
 q3

## [1] 174.5

 q1 = 106.5
 q1

## [1] 106.5

 iqr = q3 - q1
 iqr

## [1] 68

 q1 - 1.5\*iqr

## [1] 4.5

 q3 + 1.5\*iqr

## [1] 276.5

 sort(htwt$Weight)

## [1] 82 87 87 101 103 110 112 119 119 122 125 151 155 157 159 190 191 195 199
## [20] 228

Since 4.5 < 82 and 228 < 276.5, there do not appear to be any observations that appear to be potential outliers. If we change the 82 value to 8.2 (decimal place typo) and 228 to 328 (regular typo) things change a little.

 # Make a copy of the weight data
 wt = htwt$Weight
 wt

## [1] 159 155 157 125 103 122 101 82 228 199 195 110 191 151 119 119 112 87 190
## [20] 87

 # Change the 82 to 8.2
 wt[8] = 8.2
 wt[9] = 328
 wt

## [1] 159.0 155.0 157.0 125.0 103.0 122.0 101.0 8.2 328.0 199.0 195.0 110.0
## [13] 191.0 151.0 119.0 119.0 112.0 87.0 190.0 87.0

 # The quartiles are unchanged
 q3 = 174.5
 q3

## [1] 174.5

 q1 = 106.5
 q1

## [1] 106.5

 iqr = q3 - q1
 iqr

## [1] 68

 q1 - 1.5\*iqr

## [1] 4.5

 q3 + 1.5\*iqr

## [1] 276.5

 sort(wt)

## [1] 8.2 87.0 87.0 101.0 103.0 110.0 112.0 119.0 119.0 122.0 125.0 151.0
## [13] 155.0 157.0 159.0 190.0 191.0 195.0 199.0 328.0

Now we see that 4.5 < 8.2, so the lower adjacent value is unchanged. However, 199 < 276.5 < 328, so the upper adjacent value is set to 199 and we view 328 as a possible outlier.

 bwplot(~wt)



## Side-by-Side Boxplots

If we want to compare the data between groups, it is possible to create side-by-side boxplots. For the ***htwt*** data we have the factor variable ***Group*** which represents the sex of the individual. Boxplots comparing the weights of the sexes are computed below.

 # Lattice
 bwplot(Group~Weight, data=htwt)



 # ggplot2
 ggplot(htwt, aes(Group,Weight)) + geom\_boxplot() + xlab("Sex")



Note that the median of the females is less than the lower quartile of the males. Similarly, the median of the males is greater than the upper quartile of the females. So, it appears that males are typically heavier than females. We also note that there appears to be more variability in the males — the IQR of the males is greater than that of the females.