av.var.boxplots.chpt2

Astudent

9/6/2023

```
#click on this-
#Note:you may need to import some packages but the program will tell you  if so.
#these libraries are needed for filter and more advanced plots
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
#replace this with your load statements
load("~/Desktop/Rnotes/census1880midlancpa.rda")
load("~/Desktop/Rnotes/titanic_tr.rda")
load("~/Desktop/Rnotes/unicefbas2016.rda")
```

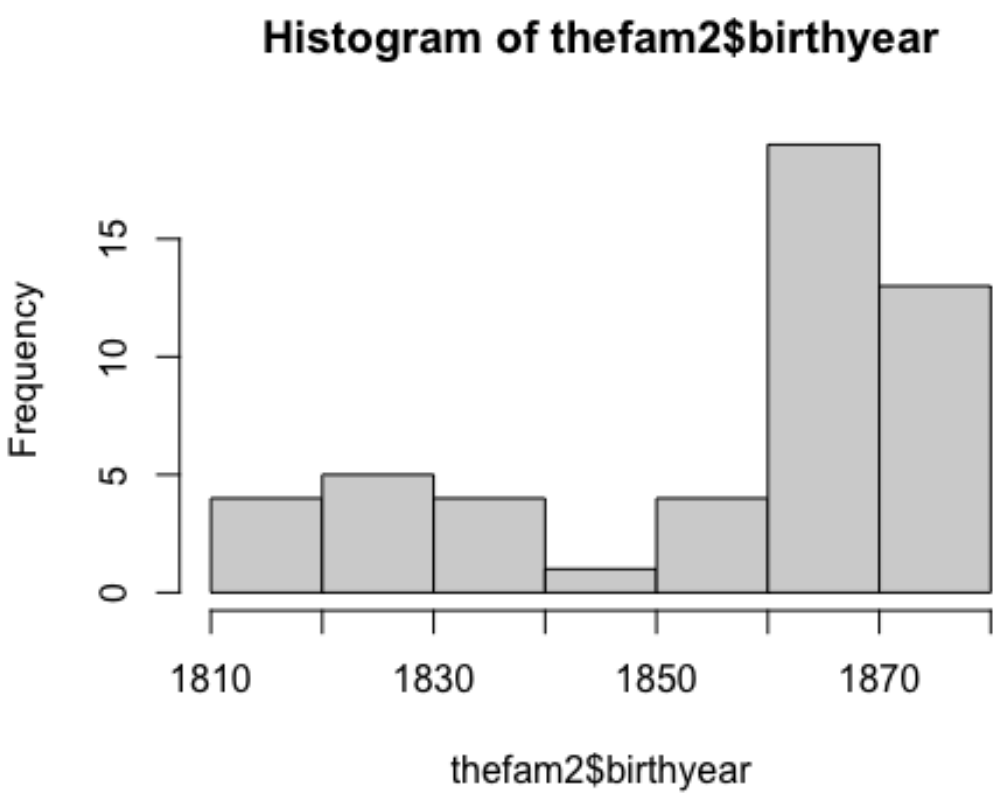
BASIC STATS-chapter 2

A.Finding the center: data: x1,x2,…,xn 1)mean=( )/_____ (student fill in) If you have two numbers the mean of them is the midpoint of the line segment between them. 1,7,10: mean is ?

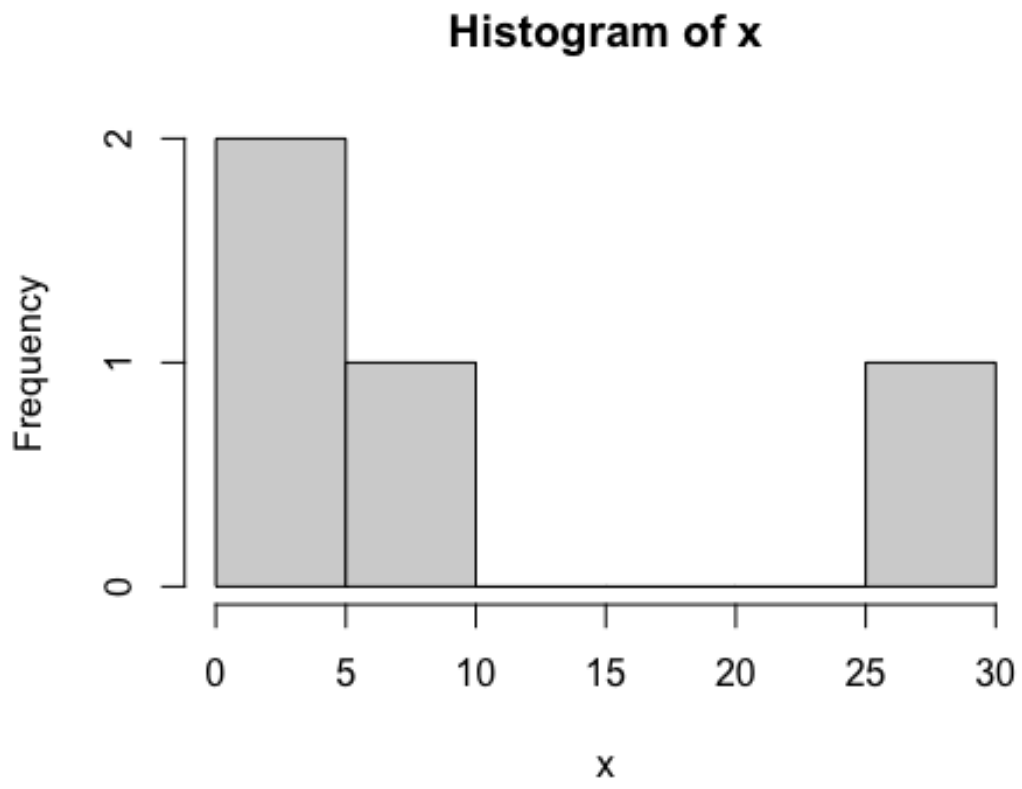
```
v=c(6, 7, 15, 36, 39, 40, 41, 42, 43, 47, 49)
mean(v)
```

```
## [1] 33.18182
```

```
hist(thefam2$birthyear)
```




```
mean(thefam2$birthyear)

## [1] 1856.76

x=c(1,3,10,30)
hist(x,breaks=10)
```

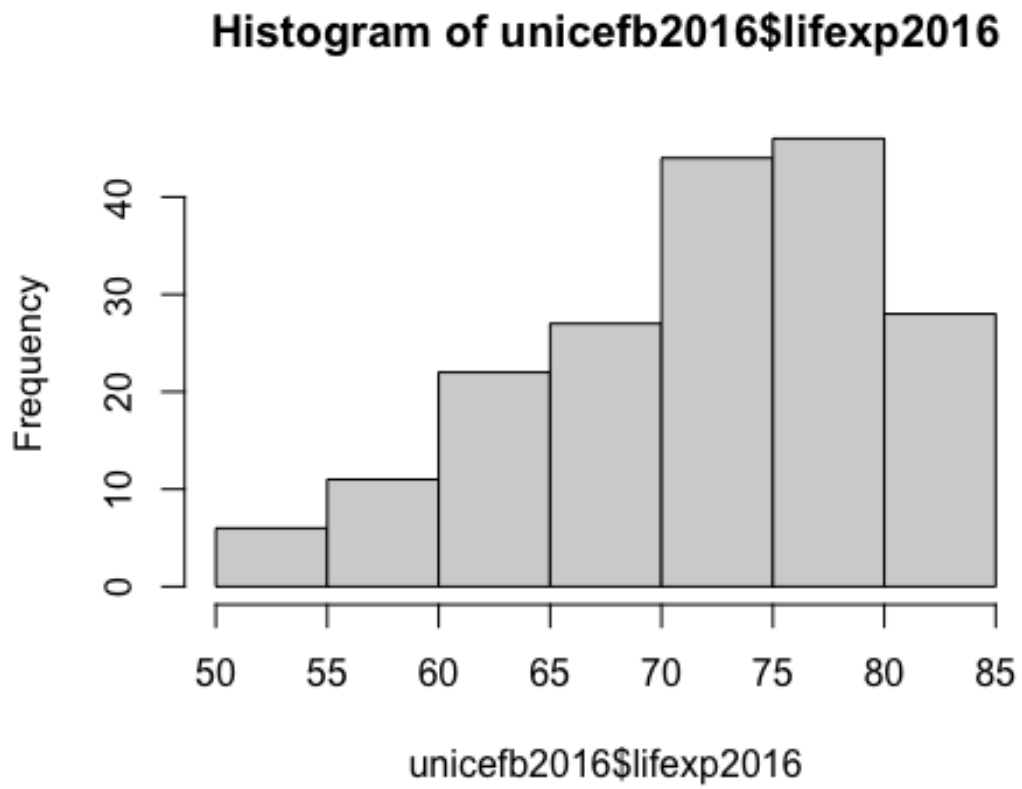



```
mean(x) # how many below this? Why not 2?
```

```
## [1] 11
```

In the following there is missing data-NA's-so we must remove them in calculating using na.rm=TRUE

```
hist(unicefb2016$lifexp2016)
```




```
mean(unicefb2016$lifexp2016, na.rm=TRUE)

## [1] 71.61411

#—Does that seem like the center?
```

2) Another form of center : median: half the data is the left half to the right.The median of 1,3,10,30 is (3+10)/2=6.5 The median of an odd number of entries is the middle value one (after they are ordered); of an even number is the mean of the middle two.

If the data is skewed, the median is a better measure of the center.

If the data is symmetric, the mean is – (student fill in) the median.

If the data is skewed right, the mean is – (student fill in) the median.

R code for median is median.

In each of the above snippets add code to compute the medians. Based on the histograms, do the results fit with your expectations?

B.Spread of the data: data: x1,x2,…,xn
xbar=mean 1)variation=( )/_____ If you have two numbers the variation of them is

```r
x=c(2,8)
var(x)
```

```
## [1] 18
```

```r
  (6^2)/2
```

```
## [1] 18
```

```r
#why are these the same?
```

standard deviation = square root of the variation sd(x) Add code to the above snippets to compute the standard deviation.

2) Another measure of spread: five number summary: minimum, Q1(first quartile), median, Q3(third quartile), maximum. Quartiles divide the data into quarters. 25% of your data is at or below the first quartile.25% is above Q3, the third quartile. Just as we had to be careful about the median as to whether there is a middle number in the data, we have to fuss a bit with quartiles and there are differing opinions on how to do that.There is only a difference if there is an odd number of data points.

The explanation below is copied from Wikipedia. For an exam, project or quiz use code from R for Q1 and Q3.For the online homework you will have to use the books definition.

Method 1 (Used by the book) Use the median to divide the ordered data set into two halves. If there is an odd number of data points in the original ordered data set, do not include the median (the central value in the ordered list) in either half. If there is an even number of data points in the original ordered data set, split this data set exactly in half. The lower quartile value is the median of the lower half of the data. The upper quartile value is the median of the upper half of the data. This rule is employed by the TI-83 calculator boxplot and "1-Var Stats" functions.

example Q1 for (1,2,3,4,5) is 1.5, Q3 is 4.5

Method 2 (used by R's command fivenum) Use the median to divide the ordered data set into two halves. If there are an odd number of data points in the original ordered data set, include the median (the central value in the ordered list) in both halves. If there are an even number of data points in the original ordered data set, split this data set exactly in half. The lower quartile value is the median of the lower half of the data. The upper quartile value is the median of the upper half of the data. The values found by this method are also known as "Tukey's hinges"

example Q1 for (1,2,3,4,5) is 2, Q3 is 4

```r
x=c(1,2,3,4,5)
fivenum(x) #gives the min, Q1, median, Q3, max
```

```
## [1] 1 2 3 4 5
```

.

If you want to use R for homework for Q1 and Q3, when there is an odd number, remove the middle number.See below.

```
x=c(11,2,32,4,5,12,6)
fivenum(x)
```

```
## [1]  2.0  4.5  6.0 11.5 32.0
```

```
y=sort(x)
```

```
y
```

```
## [1]  2  4  5  6 11 12 32
```

```
#find length y
length(y)
```

```
## [1] 7
```

```
#check if even or odd
length(y)%%2
```

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

```
midindex=(length(y)+1)/2
midindex #index of middle one
```

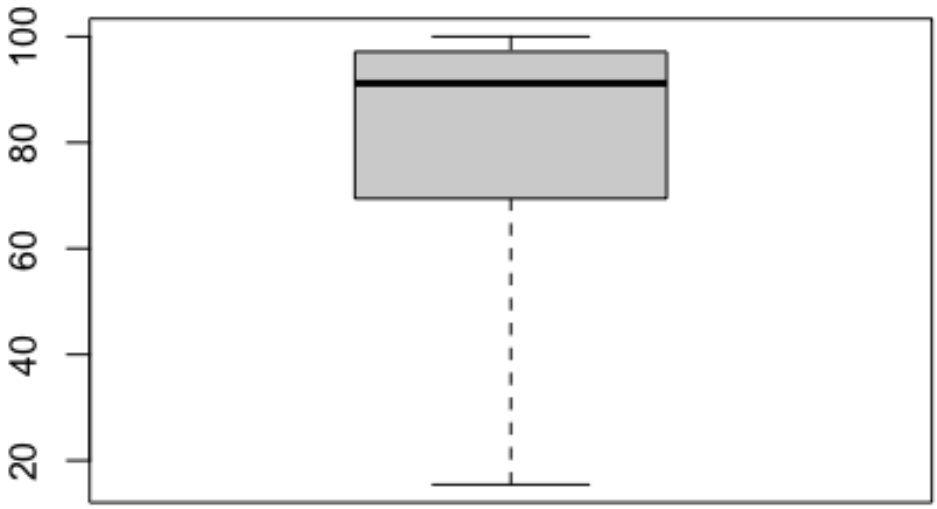
```
## [1] 4
```

```
#remove middle one:
y1=y[-midindex]
fivenum(y1)
```

```
## [1]  2  4  8 12 32
```

Boxplot: box with top Q3, bottom Q1, and median also marked and whiskers to the lower and upper data points. This is very useful for helping visualize and compare distributions of data.

```
boxplot(unicefb2016$litrt,range=0)
```




```
#What do you note with this? Where is the middle 50% of the countries'
literacy rates? Where is the lowest 25%?
```

Interquartile range: IQR=Q3-Q1

```
f<-fivenum(unicefb2016$litrt,na.rm=TRUE)
f

## [1]  15.45670  69.42539  91.18136  97.12875 100.00000

Iqr=IQR(unicefb2016$litrt,na.rm=TRUE)
Iqr

## [1] 27.70336

f[4]-f[2]

## [1] 27.70336
```

An outlier is defined as any data point which is more than 1.5*IQR above Q3 or below Q1.

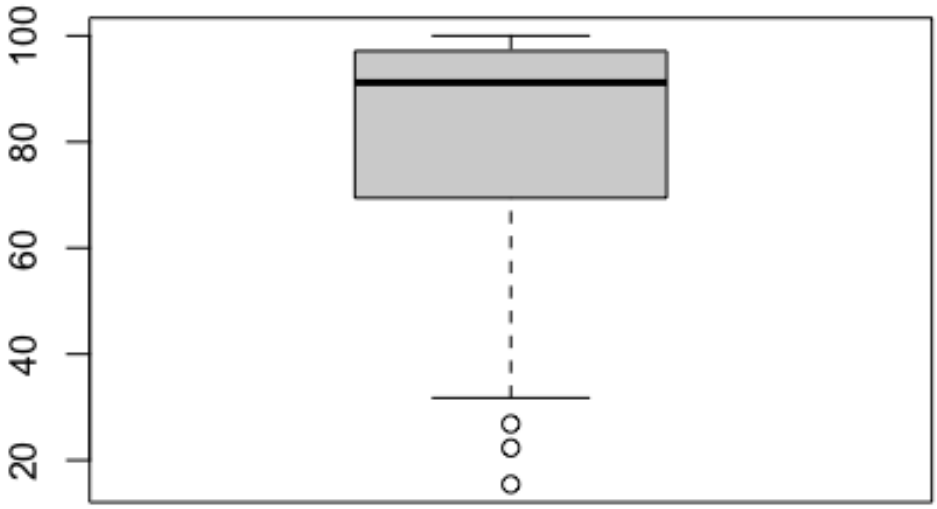
```
f[4]+1.5*Iqr

## [1] 138.6838
```

```
f[2]-1.5*Iqr
```

```
## [1] 27.87035
```

Have some low outliers. How many? We can get the boxplot to show outliers. Just omit the parameter range =0.

```
boxplot(unicefb2016$litrt)
```

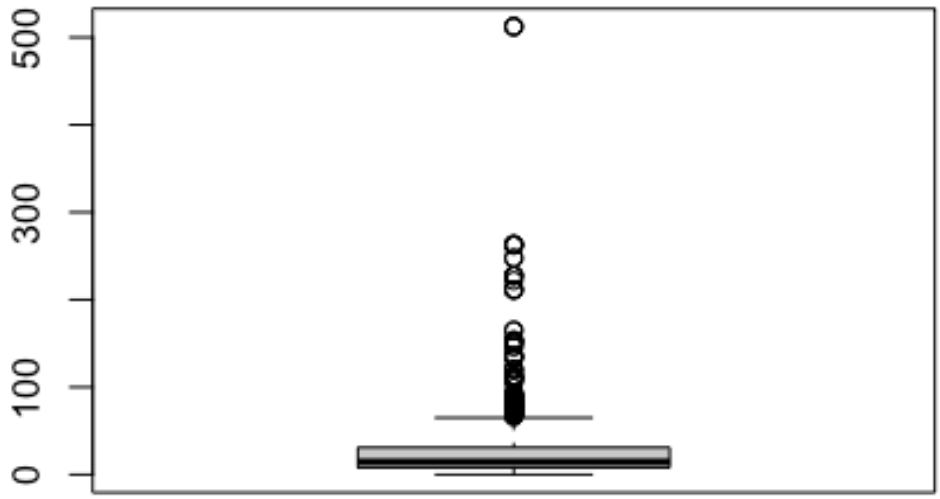



```
#Lower whisker is now at Q1-1.5*IQR
#three outliers
#What countries are these?
lowlitrtind=which(unicefb2016$litrt<f[2]-1.5*Iqr&!is.na(unicefb2016$litrt))
#tricky point of getting rid of na's in ltrt by adding the condition
!is.na(unicefb2016$litrt) using & (and)
unicefb2016$countries.and.areas[lowlitrtind]
```

```
## [1] "Chad"        "Niger"       "South Sudan"
```

Draw a boxplot of the fares of titanic passengers

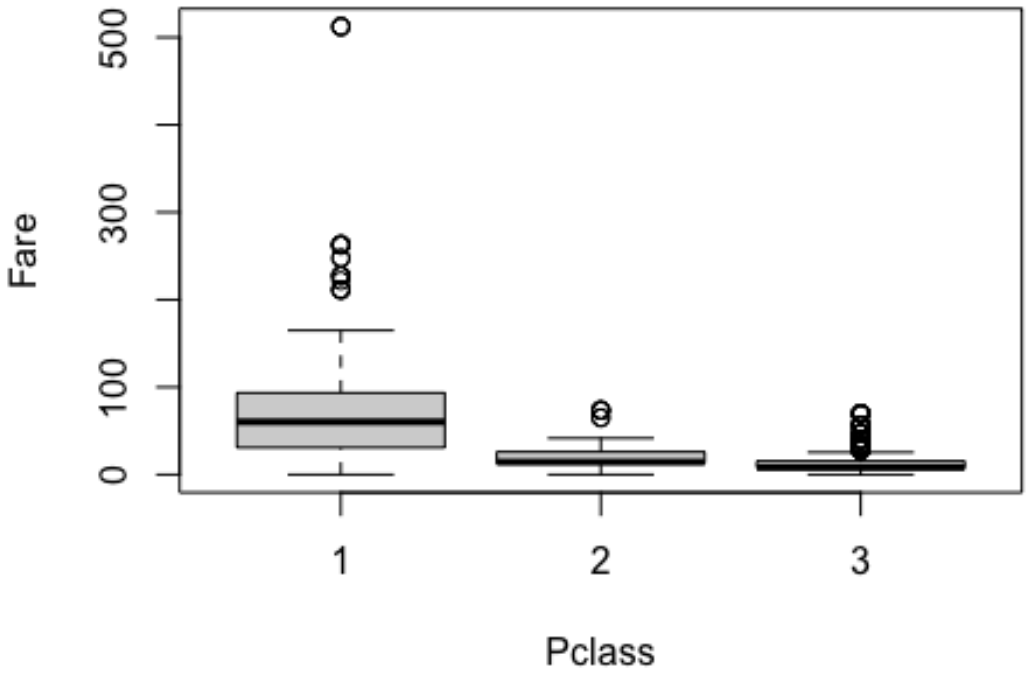
```
boxplot(titanic_train$Fare)
```




```
#What is of interest here?
```

Where boxplots become very useful is in comparing groups in the data.Let's look at fares according to Pclass

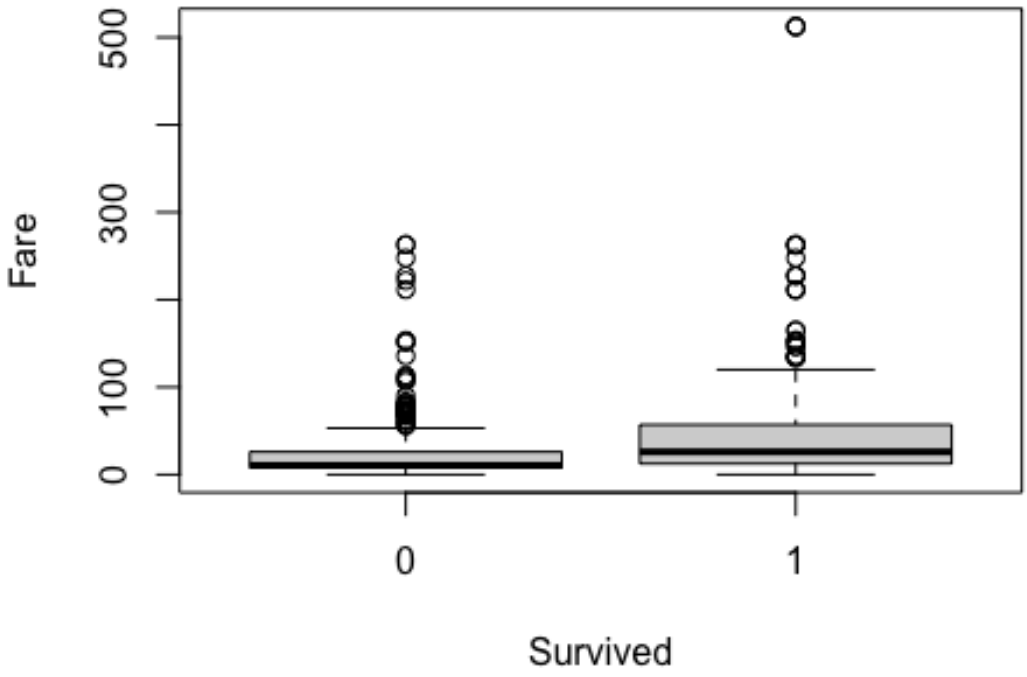
```
boxplot(Fare~Pclass,data=titanic_train)
```




```
#Note the format: variable~group variable, data=dataset
```

Let's look at fares according to survival

```
boxplot(Fare~Survived,data=titanic_train)
```



Look at u51990 in the unicef data grouped according to classf. Look at litrt in the unicef data grouped according to classf.

Look at litrt of those countries whose under five 2016 rate is below 60.

```
below60foru52016=unicefb2016%>%filter(u52016<60&!is.na(u52016)&is.na(litrt))
#compute/draw and describe boxplot, five number summary,histogram, mean, sd
```

Homework examples:

2.33 More on Nintendo and laparoscopic surgery. In Exercise 1.38 (page 42), you examined the improvement in times to complete a virtual gall bladder removal for those with and without four weeks of Nintendo WiiTM training. The most common methods for formal comparison of two groups use $\bar{x}$ and s to summarize the data.

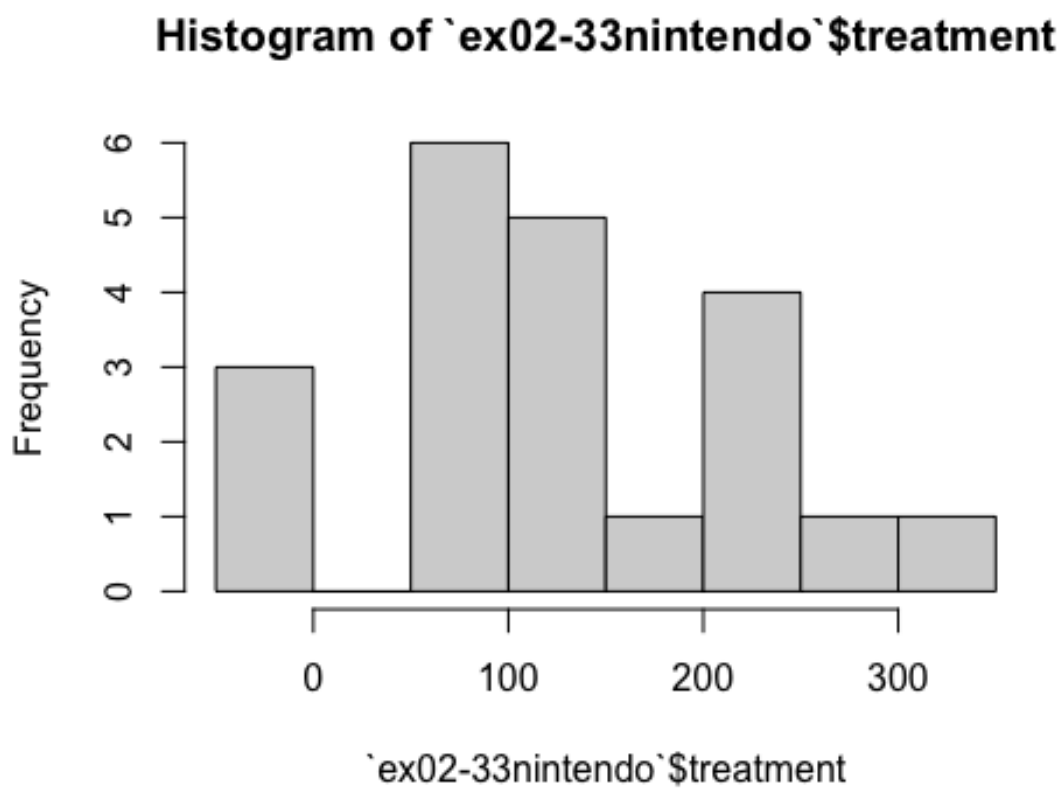
What kinds of distributions are best summarized by $\bar{x}$ and s? Do you think these summary measures are appropriate in this case? In the control group, one subject improved his/her time by 229 seconds. How much does removing this observation change $\bar{x}$ and s for the control group? You will need to compute $\bar{x}$ and s for the control group, both with and without the high outlier. Compute the median for the control group with and without the high outlier. What does this show about the resistance of the median and $\bar{x}$?

```
load("~/Desktop/Rnotes/R/chapter_2/ex02-33nintendo.rda")
```

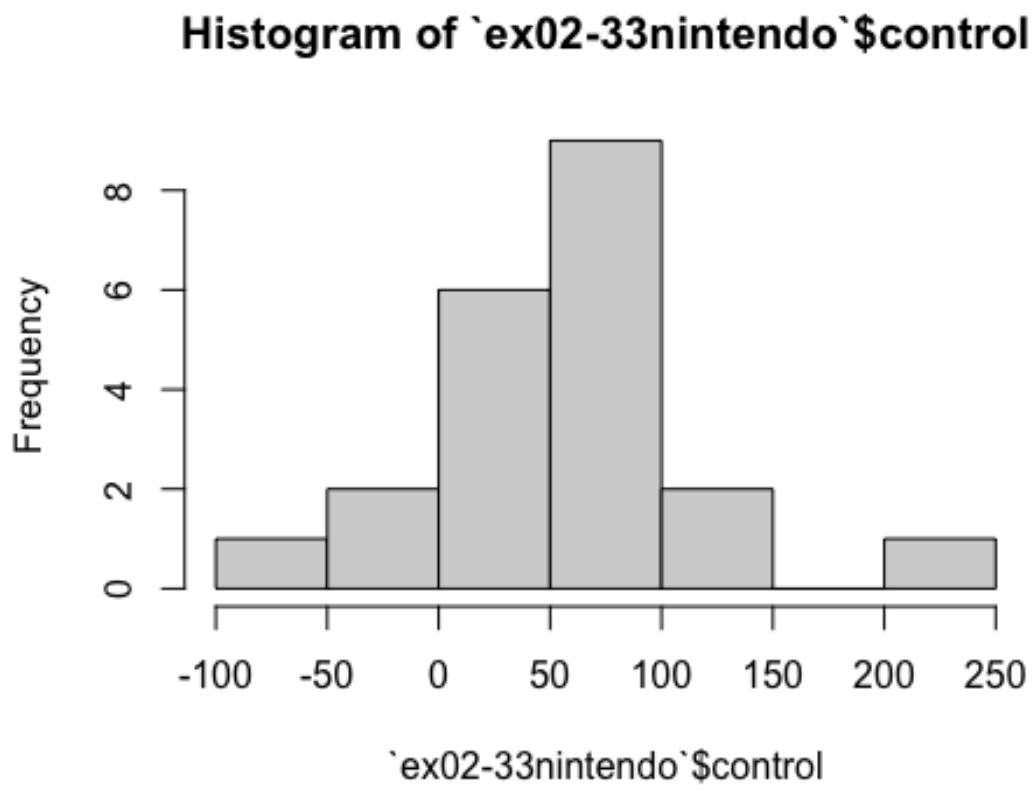
```
str(`ex02-33nintendo`)
```

```
## 'data.frame':    21 obs. of  2 variables:
##  $ treatment: int  281 134 186 128 84 243 212 121 134 221 ...
##  $ control  : int  21 66 54 85 229 92 43 27 77 -29 ...
```

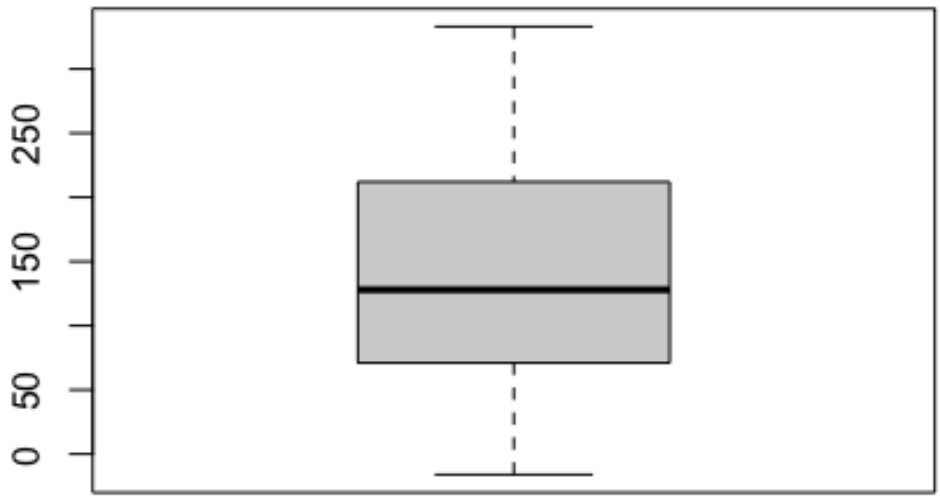
```
hist(`ex02-33nintendo`$treatment)
```



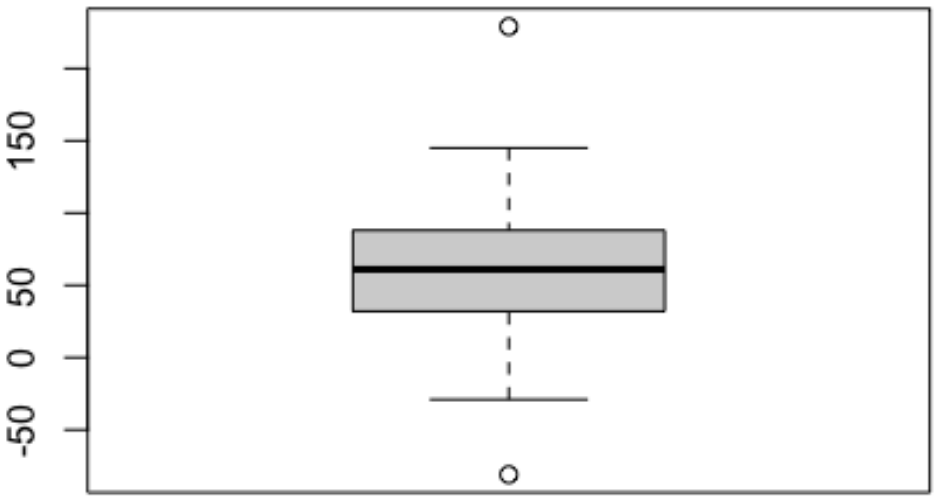

```
hist(`ex02-33nintendo`$control)
```




```
boxplot(`ex02-33nintendo`$treatment)
```




```
boxplot(`ex02-33nintendo`$control)
```




```
ex2.33.rmvhigh.otlr=`ex02-33nintendo`%>%filter(control<200)
mntr=mean(`ex02-33nintendo`$treatment)
mnct=mean(`ex02-33nintendo`$control)
mdtr=median(`ex02-33nintendo`$treatment)
mdct=median(`ex02-33nintendo`$control)
mntr2=mean(ex2.33.rmvhigh.otlr$treatment)
mnct2=mean(ex2.33.rmvhigh.otlr$control)
mdtr2=median(ex2.33.rmvhigh.otlr$treatment)
mdct2=median(ex2.33.rmvhigh.otlr$control)
mntr; mntr2
```

```
## [1] 132.2381
```

```
## [1] 134.65
```

```
mnct; mnct2 #8.5 difference
```

```
## [1] 59.71429
```

```
## [1] 51.25
```

```
mdtr; mdtr2
```

```
## [1] 128
```

```
## [1] 131
```

```
mdct;mdct2 #3.5 difference
```

```
## [1] 61
```

```
## [1] 57.5
```

Note that the control becomes skewed left when we remove the upper outlier. This is reflected in a lower mean than median.

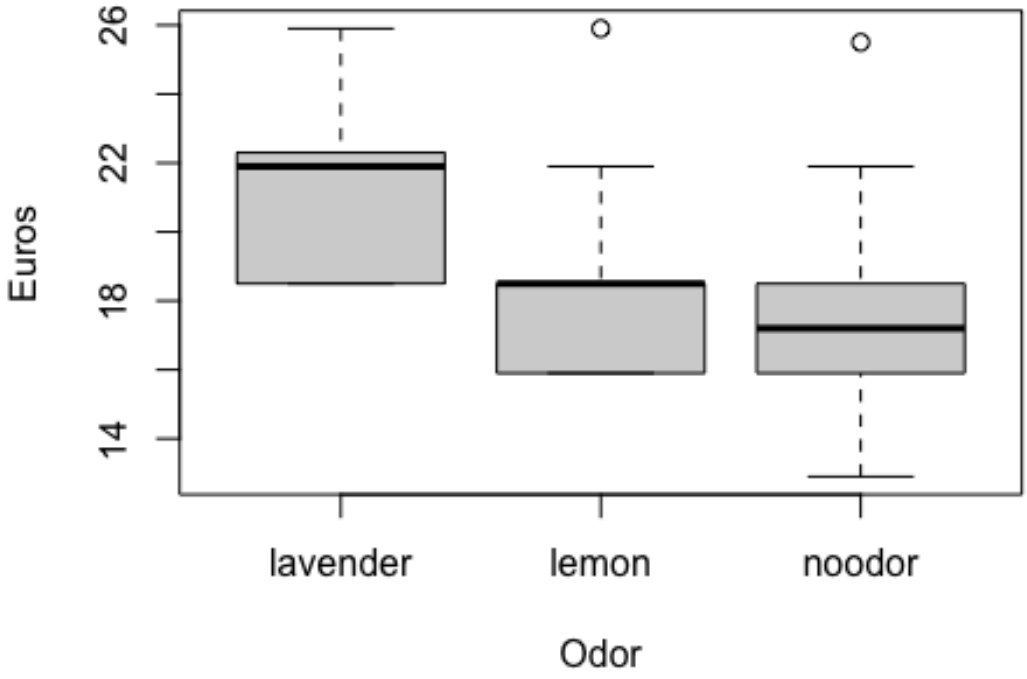
Ex 2.46. Compare the amount spent by restaurant customers in different scented environments.(no odor, lavendar, lemon). Did either smell bring better business?

```
load("~/Desktop/Rnotes/R/chapter_2/ex02-46odors.rda")
```

```
str(`ex02-46odors`)
```

```
## 'data.frame':    88 obs. of  2 variables:
##  $ Odor : Factor w/ 3 levels "lavender","lemon",..: 3 3 3 3 3 3 3 3 3 3
...
##  $ Euros: num  15.9 18.5 15.9 18.5 18.5 21.9 15.9 15.9 15.9 15.9 ...
```

```
boxplot(Euros~Odor,data=`ex02-46odors`)
```



Note the reason there are no lower whiskers for lavender and lemon is that min =Q1 for these two.

```
sort((`ex02-46odors`%>%filter(Odor=='lavender'))$Euros)
```

```
##  [1] 18.5 18.5 18.5 18.5 18.5 18.5 18.5 18.5 18.5 18.5 18.5 20.7 21.5 21.5
21.9
## [16] 21.9 21.9 21.9 21.9 21.9 21.9 21.9 22.3 22.5 22.5 22.8 24.9 24.9 25.5
25.9
```