MATH10282 Introduction to Statistics
Semester 2, 2019/2020
Examples 3, Solutions

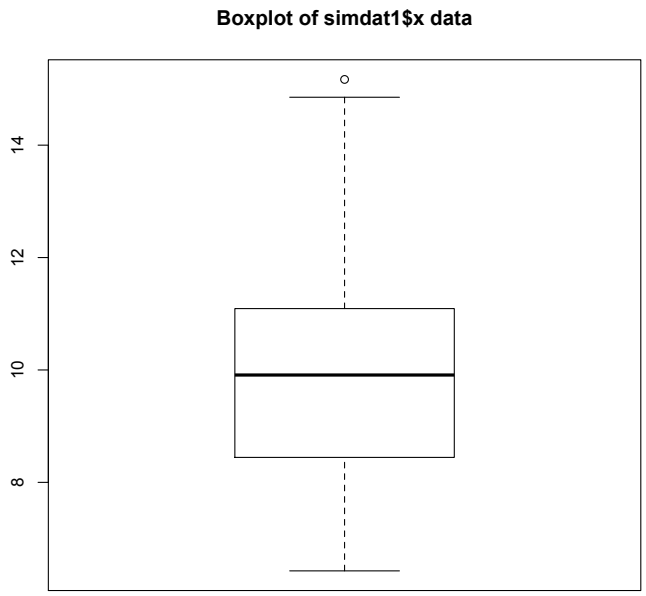
(i) Simulated data in file 'simdat1.txt'.

```
> simdat1<-read.table(file="https://minerva.it.manchester.ac.uk/~saralees/simdat1.txt",
                      header=T)
>
> names(simdat1)
[1] "case" "x"
> xs<-simdat1$x
> fns.xs<-fivenum(xs)
> fns.xs
[1]  6.425094  8.444152  9.910100 11.091922 15.169234
> iqr.xs<-fns.xs[4]-fns.xs[2]
> iqr.xs
[1] 2.64777
```

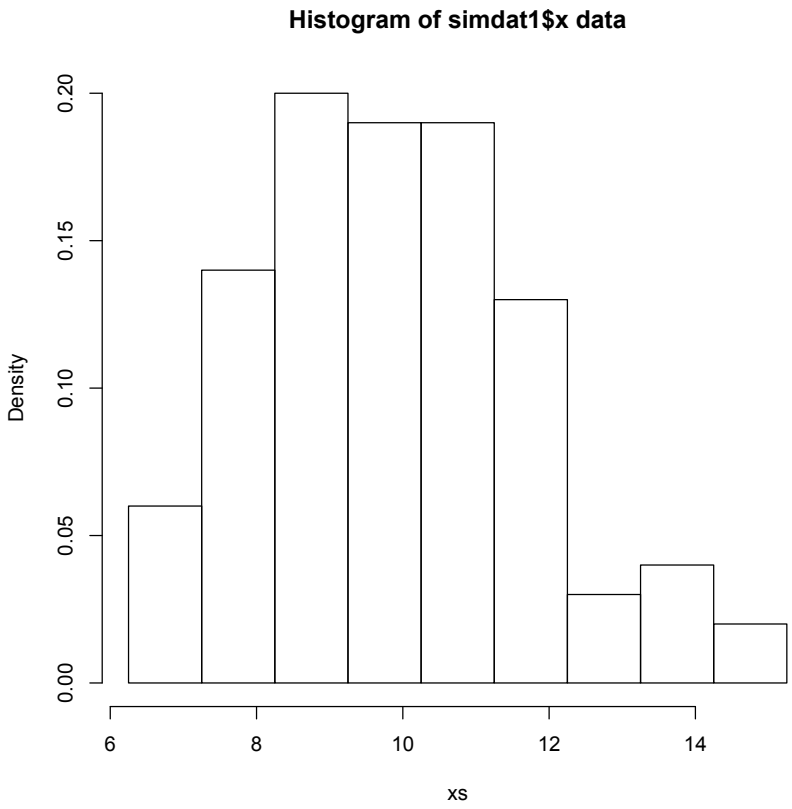
Calculate the appropriate set of differences from the five number summary and the mean to show the data is a little positively skewed.

```
> summary(xs)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  6.425   8.485   9.910   9.932  11.090  15.170
> var(xs)
[1] 3.519385
> sd(xs)
[1] 1.876002
> boxplot(xs, main="Boxplot of simdat1$x data")
```

The figure generated (see below) shows slight positive skewness and one large outlier.




```
> hist(xs, freq=F, main="Histogram of simdat1$x data")
> hist(xs, freq=F, breaks=seq(from=6, to=16.5, by=1.5),
      main="Histogram of simdat1$x data")
> hist(xs, freq=F, breaks=seq(from=6.25, to=15.25, by=1.0),
      main="Histogram of simdat1$x data")
```



The data were actually generated by computer simulation from a $N(10, 2^2)$ distribution. Sample estimates of the mean and standard deviation are 9.932 and 1.876, respectively.

(ii) Old faithful geyser data.

```
> geyser<-read.table(file="https://minerva.it.manchester.ac.uk/~saralees/geyser.txt",
                 header=T)
> names(geyser)
[1] "day"      "duration" "interval"

> xd<-geyser$duration

> summary(xd)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.700   2.300   4.000   3.576   4.400   5.200
> var(xd)
[1] 1.174948
> sd(xd)
[1] 1.08395

> boxplot(xd, main="Old Faithful geyser duration data")
```

The boxplot shows the data as appearing to be skewed - it cannot pick out the bimodality evident in the following histograms.

```
> hist(xd, main="Histogram of Old Faithful Duration Data")
> hist(xd, freq=F, breaks=16,
      main="Histogram of Old Faithful Duration Data")
> hist(xd, freq=F, breaks=seq(from=1.5, to=5.5, by=0.25),
      main="Histogram of Old Faithful Duration Data")
> hist(xd, freq=F, breaks=seq(from=1.5, to=5.5, by=0.50),
      main="Histogram of Old Faithful Duration Data")
> hist(xd, freq=F, breaks=seq(from=1.5, to=5.5, by=0.75),
      main="Histogram of Old Faithful Duration Data")
> hist(xd, freq=F, breaks=seq(from=1.5, to=5.5, by=1.00),
      main="Histogram of Old Faithful Duration Data")
```

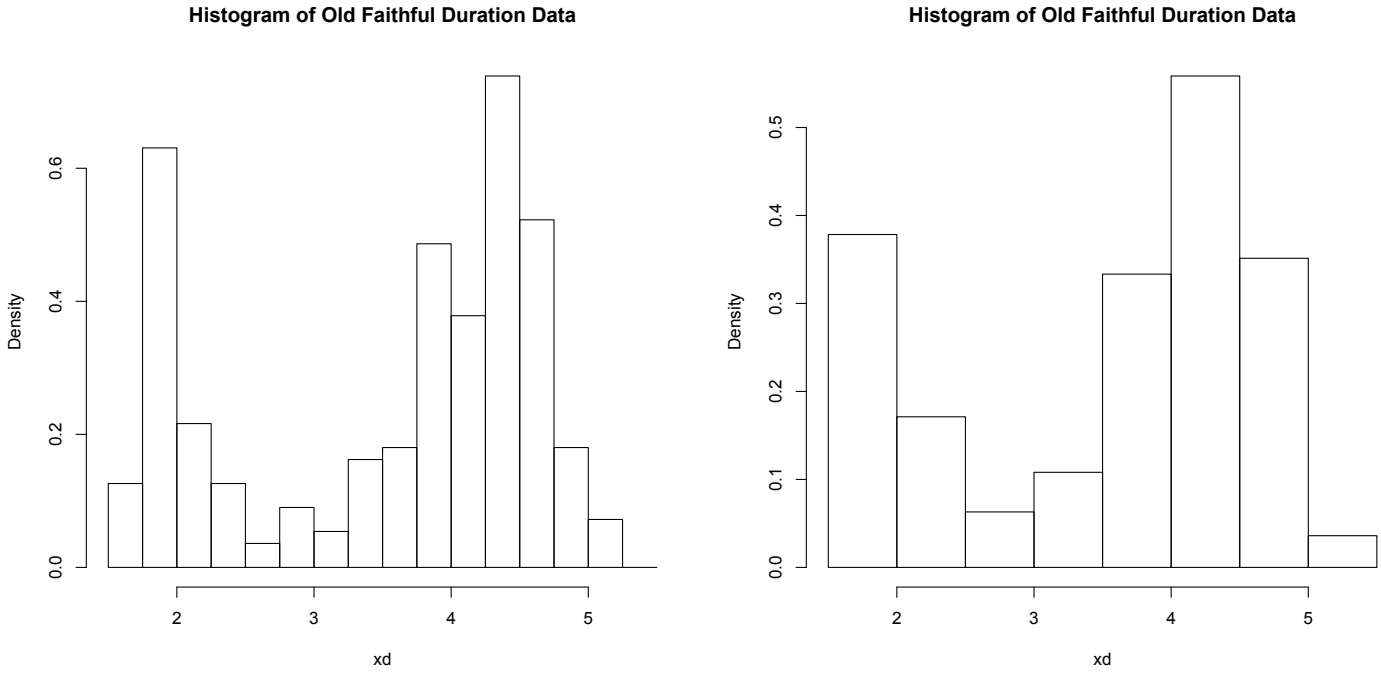


Figure: Histograms of Old Faithful geyser data
$h = 0.25$ (left) and $h = 0.5$ (right)

Use estimates of the two modes as summary measures, ie. $x = 1.875$ and $x = 4.375$. These are calculated from the histogram with `breaks=seq(from=1.5, to=5.5, by=0.25)`. This histogram is somewhat noisy as a density estimate, but it is useful for estimating the mode. The estimates of the two modes from the histogram with `breaks=seq(from=1.5, to=5.5, by=0.50)` are $x = 1.75$ and $x = 4.25$.

The full distribution appears to be a mixture of two distributions, one for eruptions of short duration and one for those of longer duration. The first is centred at approximately $x = 1.8$ and the second centred at around $x = 4.3$.

(iii) Anorexia Data.

```
> anorexia<-read.table(file="https://minerva.it.manchester.ac.uk/~saralees/anorexia.txt",
                 header=T)
> names(anorexia)
[1] "case"      "prewt"     "postwt"    "treatment"
```

```
> treat<-anorexia$treatment
> dwt<-anorexia$postwt-anorexia$prewt
> std.dwt<-dwt[treat==1]
> cog.dwt<-dwt[treat==2]
> fam.dwt<-dwt[treat==3]
> summary(std.dwt)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -12.20   -7.00   -0.35   -0.45    3.60   15.90
> summary(cog.dwt)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -9.100  -0.700   1.400   3.007   3.900  20.900
> summary(fam.dwt)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -5.300   3.900   9.000   7.265  11.400  21.500
```

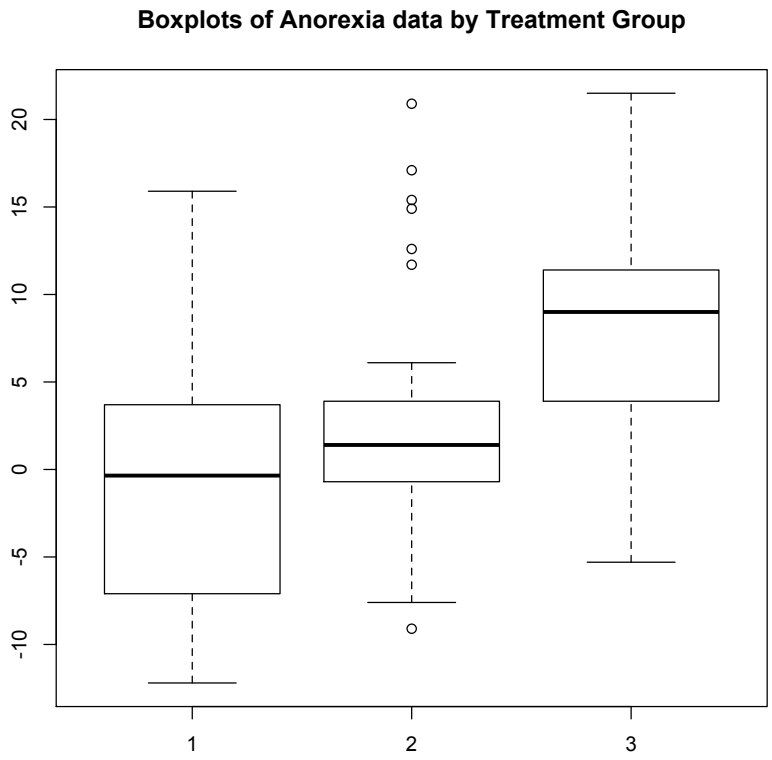
Calculate then compare and contrast the appropriate differences based on the five number summaries and means. Also, compare and contrast appropriate group summary measures e.g. means, medians, i.q.r.s and ranges. Mention any outliers in the groups.

```
> sd(std.dwt)
[1] 7.988705
> sd(cog.dwt)
[1] 7.308504
> sd(fam.dwt)
[1] 7.157421
```

The variation in each group, as measured by the standard deviations, are very similar.

```
> boxplot(dwt~treat, main="Boxplots of Anorexia
      Data by Treatment Group")
> hist(std.dwt, freq=F)
> hist(cog.dwt, freq=F)
> hist(fam.dwt, freq=F)
```

Family therapy (treat 3) was most effective, since the differences in this group have the largest mean and median. The standard treatment was least effective, as shown by the smallest mean and median.


Boxplots of Anorexia data by Treatment Group
20
15
10
5
0
-5
-10
1
2
3