Summer/Winter/Semester2 Topic2: Numerical Summaries and Boxplot 2016

Numerical Summaries
Given a sample $\{x_i\}$ and ordered data $\{x_{(i)}\}$ for $i = 1, \ldots, n$

sample mean	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
sample variance	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\Big(\sum_{i=1}^{n} x_i\Big)^2\right) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)$
sample standard deviation	s
Median (or 2nd Quartile)	$\tilde{x} = Q_2 =$ Middle data point in sorted data (for n odd) and Average of 2 middle sorted data points (for n even)
1st quartile	$Q_1 =$ Median of bottom half of sorted data
3rd quartile	$Q_3 =$ Median of top half of sorted data
Five number summary	$(x_{(1)}, Q_1, Q_2, Q_3, x_{(n)})$
Interquartile Range	$IQR = Q_3 - Q_1$
Boxplot Thresholds (for outliers)	$LT = Q_1 - 1.5IQR$, $UT = Q_3 + 1.5IQR$

Note:
(1) There are 3 formulae for the variance: the 1st one is the definition formula and the others are calculation formulae.
(2) For calculating Q_1 and Q_3, we include the median in each half set (when n is odd).

1. Australian Road Fatalities & Australian Commercial Refrigerators

```r
Road <- read.csv("http://www.maths.usyd.edu.au/u/UG/JM/MATH1005/r/StatsData/
                 AllFatalities.csv")
Age <- Road$Age
AgeN<- as.numeric(levels(Age))[Age]
class(AgeN)
fivenum(AgeN) #To get quartiles
summary(AgeN)  #To get mean
boxplot(AgeN)
```

```r
Fridge <- read.csv("http://www.maths.usyd.edu.au/u/UG/JM/MATH1005/r/StatsData/
                 Refrigerators.csv")
Efficiency <- Fridge$Efficiency..kWh.24h.m..
fivenum(Efficiency)
mean(Efficiency)
median(Efficiency)
```

For the both Age and Efficiency, what is the mean and median? Which one would you report and why?

2. Sigma Notation and Numerical Summaries

For each part, work out the answers by hand and then check in R.

(a) Given the data $x = \{1, 2, 3, 6, 7, 9\}$ and $y = \{1, 1, 2, 3, 4, 4\}$, calculate

$$\sum_{i=1}^{6} x_i \qquad \sum_{i=1}^{6} x_i^2 \qquad \sum_{i=1}^{6} x_i y_i \qquad \sum_{i=1}^{3} (x_i - 5)^2 \qquad \sum_{i=2}^{3} y_{(i)}^2$$

```
x=c(1,2,3,6,7,9)
y=c(1,1,2,3,4,4)
sum(x)
sum(x^2)
sum(x*y)
sum(((x-5)^2)[1:3])
sum((sort(y)^2)[2:3])
```

(b) Calculate the mean and standard deviation of x.

```
mean(x)
sd(x)
```

(c) If each data point in x is increased by 1, how would the mean and standard deviation change? Why? Check numerically.

```
m=x+1
mean(m)
sd(m)
```

(d) Find the quartiles of x.

```
median(x)
fivenum(x)
```

(e) (Extension: This is not examinable. Just for students who want to challenge themselves.) Given $m_i = x_i + 1$, show algebraically that $\bar{m} = \bar{x} + 1$ and $s_m^2 = s_x^2$.

3. Numerical Summaries

A sample of 36 mice was used to investigate the use of iron in Fe^+ form as a dietary supplement. The iron was given orally and was radioactively labelled so that the exact percentage of iron retained could be measured accurately. The measurements were

7.6	1.2	4.9	5.7	13.0	1.0	3.4	0.2	10.8	1.0	2.4	12.3
0.7	1.1	0.7	0.9	6.5	1.6	4.0	29.1	0.2	0.1	9.2	11.9
0.3	14.4	1.8	9.9	3.4	3.8	9.9	4.1	4.1	24.0	21.0	11.9

(a) Produce the following R output, and then use it to fill out the table.

```
x =c(7.6,1.2,4.9,5.7,13.0,1.0,3.4,0.2,10.8,1.0,2.4, 12.3,0.7,1.1, 0.7,0.9,6.5,1.6,
4.0,29.1,0.2,0.1,9.2,11.9,0.3,14.4,1.8,9.9,3.4,3.8,9.9,4.1,4.1,24.0,21.0,11.9)
length(x)
sum(x)
sum(x^2)
sort(x)
```

Size of data	Mean	Median	Standard deviation	Variance	1st Quartile	3rd Quartile	IQR

(b) What is the five number summary of x?

```
fivenum(x)
```

Note: R calculates quantiles using a few different commands. For our definition of quartiles, use the `fivenum` command. Don't use `IQR`, `summary` or `quantile`.

(c) Construct a boxplot by hand, and then check your working using R.

```
iqr=fivenum(x)[4]-fivenum(x)[2]
boxplot(x)
```

(d) In order to compare the sensitivities to outliers of the mean, median, standard deviation and IQR, the largest value is removed creating the data set $\{y\}$. Fill out the table.

```
y=c(7.6,1.2,4.9,5.7,13.0,1.0,3.4,0.2,10.8,1.0,2.4,12.3,0.7,1.1, 0.7,0.9,6.5,1.6,
    4.0,0.2,0.1,9.2,11.9,0.3,14.4,1.8,9.9,3.4,3.8,9.9,4.1,4.1,24.0,21.0,11.9)
mean(y)
median(y)
sd(y)
fivenum(y)[4]-fivenum(y)[2]
```

	Mean	Median	Std. Deviation	IQR
Data with largest value x				
Data without largest value y				
Relative Change (%)				

(e) Comment on your findings.

Extra Questions

4. Comparison of Boxplots

Students completed an online quiz consisting of 20 questions, resulting in the following marks.

Students who had Studied (A): 9 10 11 12 12 13 14 15 15 16 17 17 18
Students who had not studied (B): 1 3 5 8 9 9 10 10 12 12 14 15 16

(a) By hand, produce boxplots for A and B.

(b) Produce the boxplots in R.

```
a=c(9,10,11,12,12,13,14,15,15,16,17,17,18)
b=c(1,3,5,8,9,9,10,10,12,12,14,15,16)
boxplot(a,b)
boxplot(a,b, horizontal=TRUE,col=c("green","blue"))     # More colourful version!
```

(c) Comment on your findings.

5. Mean and median

(a) The sample average age of 5 people in a room is 30 years. A 36 year old person walks into the room. Now what is the average age of the people in the room?

(b) Suppose the median age is 30 years and a 36 year old person enters the room. Can you find the new median age from this information?

6. (Extension: Sigma Notation)

Show that the 3 formulae for variance are equal.