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	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right)$ $= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)$
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

```
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)
```

```
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?

Solution

As both data sets show some skewing, the median (robust) would be preferable, especially with Efficiency which has outliers.

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 $s_m^2 = s_x^2$.

Solution

(a)

$\sum_{i=1}^{6} x_i = 1 + 2 + \ldots 9 = 28.$

$\sum_{i=1}^{6} x_i^2 = 1^2 + 2^2 + \ldots 9^2 = 180.$

$\sum_{i=1}^{6} x_i y_i = 1 \times 1 + 2 \times 1 + \ldots 9 \times 4 = 91.$

$\sum_{i=1}^{3} (x_i - 5)^2 = (1-5)^2 + (2-5)^2 + (3-5)^2 = 29.$

$\sum_{i=2}^{3} y_{(i)}^2 = 1^2 + 2^2 = 5$, as $\{y_{(i)}\} = \{y_i\}$ as the data is already sorted.

Note: If say $\{y_i\} = \{8, 1, 4\}$, then $\{y_{(i)}\} = \{1, 4, 8\}$, so $\sum_{i=2}^{3} y_{(i)}^2 = 4^2 + 8^2 = 80$.

(b)

$$\bar{x} = \frac{1}{6}\sum_{i=1}^{6} x_i = \frac{28}{6} \approx 4.67$$

$$s = \sqrt{\frac{1}{5}\left(\sum_{i=1}^{6} x_i^2 - \frac{1}{6}\left(\sum_{i=1}^{6} x_i\right)^2\right)} = \sqrt{\frac{1}{5}\left(180 - \frac{1}{6}(28)^2\right)} \approx 3.14$$

(c)
Mean: Increases by 1 (as the full data set shifts up by 1, the average shifts up by 1).
Standard deviation: Unchanged ((as the full data set shifts up by 1, the spread is unchanged).

Check numerically:
Given $\{m_i\} = \{x_i + 1\} = \{2, 3, 4, 7, 8, 10\}$, then $\bar{m} \approx 5.67$ and $s_m^2 \approx 3.14$.

The standard deviation could also be found (the long way) in R by using the formula:

```
sqrt( 1/5*(sum(x^2) - 1/6*sum(x)^2) )
sqrt( 1/5*(sum(x^2) - 6*mean(x)^2) )
```

(d)

Sorted data is $\{x_{(i)}\} = \{1, 2, 3, 6, 7, 9\}$, so $\tilde{x} = \frac{3+6}{2} = 4.5$.
Bottom half of sorted data is: $\{1, 2, 3\}$, so $Q_1 = 2$.
Top half of sorted data is: $\{6, 7, 9\}$, so $Q_3 = 7$.

(e)
If $m_i = x_i + 1$ then:

$$\bar{m} = \frac{1}{n}\sum_{i=1}^{n}(x_i + 1) = \frac{1}{n}\sum_{i=1}^{n} x_i + \frac{1}{n}\sum_{i=1}^{n} 1 = \bar{x} + 1.$$

$$s_m^2 = \frac{1}{n-1}\sum_{i=1}^{n}(m_i - \bar{m})^2 = \frac{1}{n-1}\sum_{i=1}^{n}((x_i + 1) - (\bar{x} + 1))^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = s_x^2.$$

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.

Solution

(a)

Size of data	Mean	Median	Standard deviation	Variance	1st Quartile	3rd Quartile	IQR
36	6.61	4.05	7.09	50.26	1.05	10.35	9.3

Calculations:

$$\bar{x} = \frac{238.1}{36} = 6.61$$

$$\tilde{x} = \frac{4.0 + 4.1}{2} = 4.05$$

$$s = \sqrt{\frac{1}{35}\left(3333.85 - \frac{1}{36}(238.1)^2\right)} \approx 7.09$$

$$s^2 \approx 50.26$$

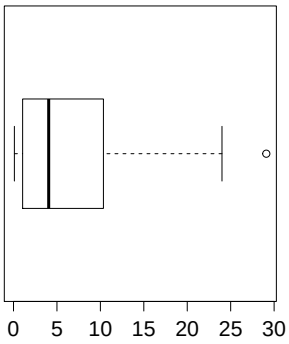
$$Q_1 = \frac{1 + 1.1}{2} = 1.05$$

$$Q_3 = \frac{9.9 + 10.8}{2} = 10.35$$

$IQR = 10.35 - 1.05 = 9.3$.

(b) 5 number summary is: $(0.1, 1.05, 4.05, 10.35, 29.1)$

(c) Construction of Boxplot:
1. Draw a box from $Q_1 = 1.05$ to $Q_3 = 10.35$, with a line within the box for $Q_2 = 4.05$.
2. Calculate $LT = Q_1 - 1.5 \times IQR = -12.9$ and $UT = Q_3 + 1.5 \times IQR = 24.3$, which determines the maximum possible length of the whiskers.
3. Draw a whisker from the box to the nearest points within LT and UT: this is 0.1 and 24.0
4. Any points lying outside the thresholds are outliers (designated by circles): 29.1.



(d)

	Mean	Median	Standard deviation	IQR
Data with largest value x	6.61	4.05	7.09	9.3
Data without largest value y	5.97	4	6.04	8.85
Relative Change (%)	(6.61-5.97)/6.61 = 9.6 %	(4.05 -4)/4.05 = 1.2 %	(7.09-6.04)/7.09 = 14.8 %	(9.3-8.85)/9.3 = 4.8 %

(e) Comment: Notice that the median and IQR have small relative change compared to the mean and sd, as they are robust.

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.

Solution

(a)

For A:
$\tilde{x} = x_{(7)} = 14$
$Q_1 = x_{(4)} = 12$
$Q_3 = x_{(10)} = 16$
$IQR = 16 - 12 = 4$
$LT = Q_1 - 1.5 * IQR = 12 - 6 = 6$
$UT = Q_3 + 1.5 * IQR = 16 + 6 = 22$
Comparing to sorted data, we see there are no data points outside LT and UT, hence there are no outliers.

For B:
$\tilde{x} = x_{(7)} = 10$
$Q_1 = x_{(4)} = 8$
$Q_3 = x_{(10)} = 12$
$IQR = 12 - 8 = 4$
$LT = Q_1 - 1.5 * IQR = 8 - 6 = 2$
$UT = Q_3 + 1.5 * IQR = 12 + 6 = 18$

Comparing to sorted data, we see there is one data point lower than LT, hence there is one outlier at 1.

Check quartiles in R:

```
fivenum(a)
fivenum(b)
```

Check your 2 boxpots against output in (b).

(c) Comments: The students who hadn't studied (B) have a lower median and a much bigger spread of marks than the students who had studied (A). One student in B has a mark (1) which is much lower than the whole rest of cohort.

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?

Solution

Let x_i denote the i^{th} person in the room. Initially there are 5 people in the room: $\bar{x} = \frac{1}{5}\sum_{i=1}^{5} x_i = 30$.

Hence the sum of the ages of the 5 people is $\sum_{i=1}^{5} x_i = 150$.

When person 6 enters the room the sum of the ages is $\sum_{i=1}^{6} x_i = 150 + 36$. Thus the new mean age will be $\bar{x} = \frac{1}{6}\sum_{i=1}^{6} x_i = 31$.

Order the individuals by age, initially we can deduce that $x_{(3)} = 30$ by the median. When person 6 enters the room the median will now be $Q_2 = \frac{x_{(3)} + x_{(4)}}{2}$, as there is no way to determine $x_{(4)}$ we can not deduce the new median age.

6. (Extension: Sigma Notation)

Show that the 3 formulae for variance are equal.

Solution

$$
\begin{aligned}
s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n}\overline{x}^2 - 2\sum_{i}^{n} x_i\overline{x}\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 + n\overline{x}^2 - 2\overline{x}\sum_{i}^{n} x_i\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 + n\overline{x}^2 - 2\overline{x}n\frac{\sum_{i}^{n} x_i}{n}\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 + n\overline{x}^2 - 2\overline{x}n\overline{x}\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 + n\overline{x}^2 - 2n\overline{x}^2\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right)
\end{aligned}
$$