Summer/Winter/Semester2 **Tutorial 3** 2015

This tutorial explores sigma notation, numerical summaries and the boxplot.

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	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)$
sample standard deviation	$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)}$
1st quartile	$Q_1 = \frac{x_{(\lceil \frac{n}{4} \rceil)} + x_{(\lfloor \frac{n}{4}+1 \rfloor)}}{2}$
2nd quartile (or median)	$Q_2 = \tilde{x} = \frac{x_{(\lceil \frac{n+1}{2} \rceil)} + x_{(\lfloor \frac{n+1}{2} \rfloor)}}{2}$
3rd quartile	$Q_3 = \frac{x_{(\lceil \frac{3n}{4} \rceil)} + x_{(\lfloor \frac{3n}{4}+1 \rfloor)}}{2}$
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: There are 2 formulae for the variance: the 1st one is the definition formula and the 2nd one is the calculation formula. The 2nd formula can also be written as $s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right)$.

1. Sigma Notation, Ceiling and Floor Functions

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

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

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

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

(d) Find the 1st and 3rd quartiles of x, by first calculating

$$x_{(\lceil \frac{n}{4} \rceil)} \qquad x_{(\lfloor \frac{n}{4}+1 \rfloor)} \qquad x_{(\lceil \frac{3n}{4})\rceil)} \qquad x_{(\lfloor \frac{3n}{4}+1 \rfloor)}$$

(e) Check your working in (a) to (d) using R.

```
> 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])
```

```
> mean(x)
> sd(x)
> quantile(x,type=2)

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

(f) 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 $s_m^2 = s_x^2$.

2. 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) Using the the following R output, fill out the table.

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

```
> sum(x)
[1] 238.1
> sum(x^2)
[1] 3333.85
> sort(x)
 [1]  0.1  0.2  0.2  0.3  0.7  0.7  0.9  1.0  1.0  1.1  1.2  1.6  1.8  2.4
[15]  3.4  3.4  3.8  4.0  4.1  4.1  4.9  5.7  6.5  7.6  9.2  9.9  9.9 10.8
[29] 11.9 11.9 12.3 13.0 14.4 21.0 24.0 29.1
```

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

(c) Construct a boxplot by hand.

(d) In order to compare the sensitivities to outliers of the mean, median, standard deviation and IQR, remove the largest value and recompute these four numerical summaries. Compute the relative change in each, as a percentage.

	Mean	Median	Standard deviation	IQR
Data without largest value				
Relative Change (%)				

Hint: the sum and sum of squares become 209 and 2487.04.

(e) Comment on your findings.

(f) Check your answers in (a) to (c) using R.

```
> x=scan(file=url("http://www.maths.usyd.edu.au/MATH1005/r/wk3q1.txt"))
```

```
> length(x)
> mean(x)
> median(x)
> sd(x)
> var(x)
> quantile(x,type=2)
> iqr=quantile(x,type=2)[4]-quantile(x,type=2)[2]
> boxplot(x)
```

3. 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) Scan the data into R and produce the boxplots.

```
>boxplot(a,b)
>boxplot(a,b, horizontal=TRUE,col=c("green","blue"))    # More colourful version!
```

(c) Comment on your findings.

4. 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?

5. Extension (Sigma Notation)

Show that the 2 formulae for variance are equal.