The University of Sydney
School of Mathematics and Statistics

Solutions to Tutorial Week 2

MATH1905: Statistics (Advanced) Semester 2, 2017

Web Page: `http://sydney.edu.au/science/maths/MATH1905`
Lecturer: Michael Stewart

1. The following data refer to the number of days with rain in July for Sydney from 2001 - 2008:

12, 2, 5, 8, 7, 13, 3, 9

Using your calculator, the average number of days with rain is closest to

(a) 12.00 (b) 7.50 (c) 7.38 (d) 3.57 (e) none of the above

Solution: (c) since $n = 8$, $\sum x_i = 59$ we have $\bar{x} = 7.375$ days.

2. The (sample) standard deviation for the number of days with rain is:

(a) 15.70 (b) 13.73 (c) 3.71 (d) 3.69 (e) 3.96

Solution: (e), since $\sum x_i^2 = 545$ we use the computing formula (see also 5(b) and 5(c))

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} = 545 - \frac{59^2}{8} = \frac{879}{8}$$

and so $s_x = \sqrt{\frac{879/8}{7}} \approx 3.96$ days.

3. The (Tukey) five number summary for this data set is:

(a) `2 4 7.5 8 9`
(b) `12 2.5 7.5 10.5 13`
(c) `2 3 8 12 13`
(d) `12 5 7 13 9`
(e) `2 4 7.5 10.5 13`

Solution: (e). The ordered data are: 2 3 5 7 8 9 12 13. Since there are $n = 8$ values (an even number) the median is the average of the two middle scores i.e. 7.5. Furthermore, the lower quartile is the median of the lower half i.e. the median of the lower 4 values, so the average of the 2nd and 3rd smallest values, i.e. 4. Similarly the upper quartile is the average of 9 and 12, i.e. 10.5.

4. Refer to the data in question 1

(a) Check your answers to questions 1, 2 and 3 using R. Read data in using a command of the form `x <- c(...)`, then use commands `mean(x)`, `sd(x)` and `fivenum(x)`.

Solution:

```
x=c(12, 2, 5, 8, 7, 13, 3, 9)
mean(x)
```

```
[1] 7.375
```

```
sd(x)
```

Copyright © 2017 The University of Sydney

```
[1] 3.961872
```

```
fivenum(x)
```

```
[1]  2.0  4.0  7.5 10.5 13.0
```

(b) Note that `quantile(x)` and `summary(x)` do not agree with `fivenum(x)`. Try also `quantile(x,type=1)`, `quantile(x,type=2)`,..., `quantile(x,type=9)`, just to see how many *slightly different* methods exist for computing quartiles (and why we prefer using `fivenum(x)`!)

Solution:

```
quantile(x)
```

```
  0%   25%   50%   75%  100%
2.00  4.50  7.50  9.75 13.00
```

```
summary(x)
```

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
2.000   4.500   7.500   7.375   9.750  13.000
```

```
quantile(x,type=1)
```

```
0%  25%  50%  75% 100%
 2    3    7    9   13
```

```
quantile(x,type=2)
```

```
 0%  25%  50%  75% 100%
2.0  4.0  7.5 10.5 13.0
```

```
quantile(x,type=3)
```

```
0%  25%  50%  75% 100%
 2    3    7    9   13
```

```
quantile(x,type=4)
```

```
0%  25%  50%  75% 100%
 2    3    7    9   13
```

```
quantile(x,type=5)
```

```
 0%  25%  50%  75% 100%
2.0  4.0  7.5 10.5 13.0
```

```
quantile(x,type=6)
```

```
  0%   25%   50%   75%  100%
2.00  3.50  7.50 11.25 13.00
```

```
quantile(x,type=7)
```

```
  0%   25%   50%   75%  100%
2.00  4.50  7.50  9.75 13.00
```

```
quantile(x,type=8)
```

```
      0%       25%       50%       75%      100%
2.000000  3.833333  7.500000 10.750000 13.000000
```

```
quantile(x,type=9)
```

```
    0%     25%     50%     75%    100%
2.0000  3.8750  7.5000 10.6875 13.0000
```

(c) Suppose that in 2015 there are 20 days of rain in July. Is such an observation an outlier in the context of the data in question 1? To find out, draw a boxplot (by hand and using R) for the new data set 12, 2, 5, 8, 7, 13, 3, 9, 20 (*to be clear, is the value 20 in this dataset of 9 values regarded as an outlier, according to Tukey?*)

Solution: The new dataset has 9 values, and so the median is the 5th order statistic (the 5th smallest value) i.e. 8. Tukey's rule for quartiles in this case is the median of the lower half *including the median* that is the median of the lowest 5 values, the 3rd order statistic, 5. Similarly the upper quartile is the median of the largest 5 values, i.e. 12.

```
y=c(x,20)
y
```

```
[1] 12  2  5  8  7 13  3  9 20
```

```
sort(y)
```

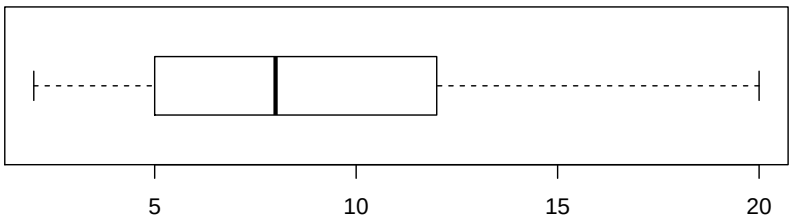
```
[1]  2  3  5  7  8  9 12 13 20
```

```
fivenum(y)
```

```
[1]  2  5  8 12 20
```

So the interquartile range (IQR) is $12 - 5 = 7$ and thus any value *more* than $1.5 \times 7 = 10.5$ away from the box is deemed an outlier, that is outside the range $[-5.5, 22.5]$. There are no (Tukey) outliers here and so the whiskers are drawn to the extreme values 2 and 20:

```
boxplot(y,horizontal=T)
```



5. Show that for any set of numbers $x_1, x_2, \ldots, x_n$ the following are true:

(a) $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$.

Solution: $\sum(x_i - \bar{x}) = \sum x_i - n\bar{x} = \sum x_i - \sum x_i = 0$.

(b) $\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$.

Solution:

$$\begin{aligned}\sum(x_i - \bar{x})^2 &= \sum(x_i^2 - 2\bar{x}x_i + \bar{x}^2) \\ &= \sum x_i^2 - 2\bar{x}\sum x_i + n\bar{x}^2 \\ &= \sum x_i^2 - 2\bar{x}(n\bar{x}) + n\bar{x}^2 \\ &= \sum x_i^2 - n\bar{x}^2\,.\end{aligned}$$

(c) If in addition we have $y_1, y_2, \ldots, y_n$ show that

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}\,.$$

Solution:

$$\begin{aligned}
S_{xy} &= \sum_{i=1}^{n}(x_i y_i - \bar{x}y_i - \bar{y}x_i + \bar{x}\bar{y}) \\
&= \sum_{i=1}^{n} x_i y_i - \bar{x}\sum_{i=1}^{n} y_i - \bar{y}\sum_{i=1}^{n} x_i + n\bar{x}\bar{y} \\
&= \sum_{i=1}^{n} x_i y_i - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)\sum_{i=1}^{n} y_i - \left(\frac{\sum_{i=1}^{n} y_i}{n}\right)\sum_{i=1}^{n} x_i + n\left(\frac{\sum_{i=1}^{n} x_i}{n}\right)\left(\frac{\sum_{i=1}^{n} y_i}{n}\right) \\
&= \sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n} - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n} + \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n} \\
&= \sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}
\end{aligned}$$

6. Compute the intercept and slope of the least-squares line associated with the following ordered pairs:

x_i:	5	3	10	1
y_i:	2	1	5	0

Your answers to 5(b) and 5(c) may be useful (see also lecture 4).

Solution: It is straightforward to compute the summary statistics

$$\sum_i x_i = 19 \qquad \sum_i y_i = 8$$
$$\sum_i x_i^2 = 135 \qquad \sum_i y_i^2 = 30$$
$$\sum_i x_i y_i = 63 \qquad n = 4$$

Then

$$S_{xx} = 135 - \frac{19^2}{4} = 179/4$$
$$S_{xy} = 63 - \frac{19 \times 8}{4} = 25$$

and so the least-squares slope is

$$b = S_{xy}/S_{xx} = \frac{4 \times 25}{179} = \frac{100}{179} \approx -0.5587$$

and the least-squares intercept is

$$a = \bar{y} - b\bar{x} = \frac{8}{4} - \frac{100 \times 19}{179 \times 4} = -\frac{117}{179} \approx -0.6536\,.$$

7. Check your answer to Q6 using R:

(a) Enter these points into R (as two separate vectors, say `x` and `y`).

(b) Create a scatterplot (with the x_i's on the horizontal axis).

(c) Use `lm()` to compute the least-squares line (see the daily min/max temperature example in lecture 4).

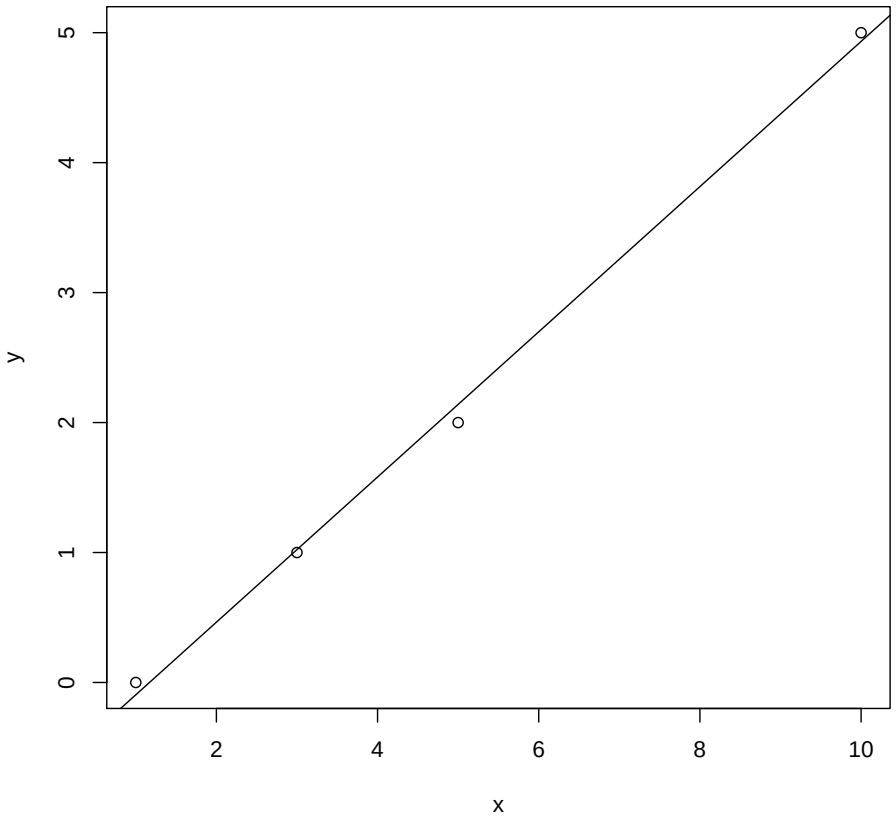
(d) Add the line to the plot.

Solution:

```
x=c(5,3,10,1)
y=c(2,1,5,0)
fit=lm(y~x)
coef(fit)
```

```
(Intercept)           x
 -0.6536313   0.5586592
```

```
plot(x,y)
abline(fit)
```



The remaining questions are provided for extra practice after the tutorial.

8. In a survey report the number of children per household was summarised using the following table.

Number of Children	Number of Households
0	7
1	4
2	8
3	4
4	2

(a) How many households were involved in the survey?

Solution: This is a frequency table. The "sample size" is just the total frequency $7+4+8+4+2=25$.

(b) Calculate the average number of children per household and the standard deviation for the data set.

Solution: To compute the average of the *original data* (as opposed to the summary in the frequency table) we need the sum of the original data. In this case the sum is

$$(0\times 7)+(1\times 4)+(2\times 8)+(3\times 4)+(4\times 2)=0+4+16+12+8=40\,.$$

So the average is $\frac{40}{25}=1.6$. The (sample) variance of the original data can be computed using the formula

$$s^2=\frac{1}{n-1}\left[\sum_{i=1}^{n}x_i^2-\frac{\left(\sum_{i=1}^{n}x_i\right)^2}{n}\right].$$

We have the sample size n and the sum $\sum_{i=1}^{n}x_i$. The sum of squares is

$$(0^2\times 7)+(1^2\times 4)+(2^2\times 8)+(3^2\times 4)+(4^2\times 2)=0+4+32+36+32=104\,.$$

therefore the sample variance is

$$\frac{1}{24}\left[104-\frac{40^2}{25}\right]=\frac{104\times 25-40^2}{24\times 25}=\frac{2600-1600}{600}=\frac{5}{3}=1.\dot{6}\,.$$

The sample standard deviation is therefore $\sqrt{\frac{5}{3}}\approx 1.291$.

(c) If there were exactly 2 adults in each household as well as the children reported above calculate the standard deviation for the total household size. Comment.

Solution: The standard deviation remains unchanged from the answer in the previous part if we simply add 2 to each data value (which is what is being done here).

9. The following list gives the number of days with rain from 1977 - 1990 for Wollongong for July, August and December.

July	2	8	6	7	6	12	8	15	7	9	11	6	12	12
August	5	7	4	4	7	3	12	6	10	9	16	10	9	9
December	8	19	7	12	13	12	18	10	16	9	16	19	15	13

Use R or do the following by hand:

(a) Provide for each month the five number summary.

Solution:

```
july=c(2,8,6,7,6,12,8,15,7,9,11,6,12,12)
fivenum(july)
```

```
[1]  2  6  8 12 15
```

Note that there are

```
length(july)
```

```
[1] 14
```

observations, so the median is the average of the two middle (i.e. 7th and 8th) order statistics, while the lower (upper) quartile is the median of the lower (upper) half of the data (i.e. of the lower(upper) seven order statistics), which is the 4th smallest (largest) observation.

```
sort(july)
```

```
[1]  2  6  6  6  7  7  8  8  9 11 12 12 12 15
```

```
sort(july)[c(7,8)]
```

```
[1] 8 8
```

```
mean(sort(july)[c(7,8)])
```

```
[1] 8
```

```
sort(july)[4]
```

```
[1] 6
```

```
sort(july)[11]
```

```
[1] 12
```

Similarly for the other two:

```
aug=c(5,7,4,4,7,3,12,6,10,9,16,10,9,9)
fivenum(aug)
```

```
[1]  3  5  8 10 16
```

```
dec=c(8,19,7,12,13,12,18,10,16,9,16,19,15,13)
fivenum(dec)
```

```
[1]  7 10 13 16 19
```

(b) Calculate the coefficient of correlation between the July and August figures and between the July and December figures. Comment on any difference.

Solution: *To be added after week 3.*

(c) Assume you had the number of days with rain in July of an additional year, i.e. your new July data is

$$2,\ 8,\ 6,\ 7,\ 6,\ 12,\ 8,\ 15,\ 7,\ 9,\ 11,\ 6,\ 12,\ 12,\ x_{15}$$

Determine the range of x_{15} such that this new observation would appear as a potential outlier in the boxplot (**hint**: consider the two cases where x_{15} is the minimum and the maximum).

Solution: This "new" dataset is just the July data from part (a), with x_{15} added to it. Recall that the order statistics of the original July data were

```
sort(july)
```

```
[1]  2  6  6  6  7  7  8  8  9 11 12 12 12 15
```

and that there are 14 observations there.

If we add another value x_{15} so there are now 15 observations, we have that

- the median is the 8th order statistic of the new data;
- the lower quartile is the median of the "lower half" *including the middle value* (since the sample size is odd); this will be the average of the 4th and 5th order statistics;
- similarly the upper quartile will be the average of the 4th and 5th largest observations.

Suppose firstly that the new observation is 0 (the smallest "possible" value in this context). This would give

```
sort(c(0,july))
```

```
[1]  0  2  6  6  6  7  7  8  8  9 11 12 12 12 15
```

```
fivenum(c(0,july))
```

```
[1]  0.0  6.0  8.0 11.5 15.0
```

The the lower quartile remains unchanged while the upper quartile is 11.5. The IQR is 5.5, the lower "outlier threshold" is then $6 - 1.5 \times 5.5 = -2.25$. So we cannot have a lower outlier here (since we cannot have a negative count).

Suppose now that the new observtion is 31 (the largest possible value in this context):

```
sort(c(31,july))
```

```
[1]  2  6  6  6  7  7  8  8  9 11 12 12 12 15 31
```

```
fivenum(c(31,july))
```

```
[1]  2.0  6.5  8.0 12.0 31.0
```

The upper quartile becomes 12 again while the IQR remains 5.5. The upper "outlier threshold" is then $12 + 1.5 \times 5.5 = 20.25$. So indeed if x_{15} was any of the values 21,22,...,31, it would be deemed an outlier (in the "new" dataset).

10. N.White collected data on the total ridge counts in fingerprints of corresponding fingers on the left and right hands of a sample of 15 Maiali aborigines from Western Arnhem Land. Calculate the coefficient of correlation between the left hand and right hand total ridge counts and construct a scatterplot of the data.

Compare the left and right hand data via boxplots. Calculate the standard deviations to determine if the spread of counts is similar on both hands. Is standard deviation an appropriate measure of spread to use in this case?

Left Hand	74	113	69	68	61	70	99	46	74	71	76	64	62	100	77
Right Hand	92	116	73	73	75	83	105	5	278	89	83	72	66	110	78

Solution: *To be added after week 3.*