The University of Sydney
School of Mathematics and Statistics

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

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

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`

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)`.

(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)`!)

(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?*)

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

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

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

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

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

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).

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.

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?

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

(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.

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.

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

(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).

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	52	78	89	83	72	66	110	78