THE UNIVERSITY OF SYDNEY

MATH1905 STATISTICS ADVANCED

Semester 2 **Tutorial Week 2** 2012

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

(a) Calculate the average number of days with rain.

(b) Calculate the standard deviation for the number of days with rain.

(c) Obtain the five number summary for this data set.

(d) Check your answer using R (in case quartiles are different, check whether or not the presented solution satisfies the definition of the lower and upper quartile, respectively).

(e) Assume in 2012 there are 20 days of rain in July. Is such an observation an outlier? 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.

2. In `R` Type

- `data(swiss)` to obtain the 'swiss' data set
- `attach(swiss)` to obtain the 6 variables from the data frame
- `help(swiss)` and read infos about this data set.

(a) Type `cor(swiss)` to obtain the matrix of pairwise correlations. What are the 3 most correlated variable pairs?

(b) Scatter plot: `plot(Education,Examination)`, `plot(Education,Fertility)`, `plot(Agriculture,Examination)`, `plot(Catholic,Fertility)`. Do you see any pattern? If yes does it agree with the corresponding correlation. What do we learn from this data analysis?

(c) Type `pairs(swiss)` to obtain all the paired scatter plots. Comment on the plots as well as on the pairwise correlations.

3. Show that for any set of numbers $x_1, x_2, \ldots, x_n$ the following is true: $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$.

4. Use R to evaluate the expressions below using the following data set:

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

(a) $\sum_{i=1}^{4} x_i$.

(b) $\sum_{i=1}^{4} x_i y_i$.

(c) $S_{xy} = \sum_{i=1}^{4} x_i y_i - (\sum_{i=1}^{4} x_i)(\sum_{i=1}^{4} y_i)/4$.

(d) $\sum_{i=1}^{4}(4x_i - 1)$.

5. 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

Assignment 1 for MATH1905 STATISTICS (due on Tuesday, 23 August, in week 5) will consist of selected questions from the Problem Sheets for weeks 1, 2, 3, 4.

1. 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.

2. 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.

Extra questions to try: *A Primer of Statistics:* Ch I page 34 Q7, 11-18 and 24.