FAMILY NAME: . . . .
OTHER NAME(S): . . . .
STUDENT NUMBER: . . . .
SIGNATURE: . . . .

THE UNIVERSITY OF NEW SOUTH WALES

SCHOOL OF MATHEMATICS AND STATISTICS

November 2013

MATH2859/MATH2099

All answers must be written in ink. Except where they are expressly required pencils may only be used for drawing, sketching or graphical work.

Probability, Statistics and Information

Answer this question in a separate book marked Question 1

1. [20 marks]

a) An engineer is interested in testing the bias in a pH meter. Data are collected on a neutral substance (pH=7.0). A sample of the measurements were taken with the data as follows:

```
6.85 7.01 6.98 7.22 6.94 6.77 7.03 7.05 6.93 7.09
```

The sample has mean of $\bar{x} = 6.987$ and standard deviation of $s = 0.1259$.

i) Determine the 5 number summary for the data.

ii) Determine whether there are any outliers. Justify your answer.

iii) Construct a boxplot for this data and comment on its features.

iv) Determine a 99% confidence interval for the true (population) mean pH measurement.

v) State any assumptions you make to determine this confidence interval. Explain whether you have sufficient information to justify making these assumptions here.

b) Engineers at a large automobile manufacturing company are trying to decide whether to purchase brand A or brand B tires for the company's new models. To help them arrive at a decision, an experiment is conducted using 12 of each brand. The tires are run until they wear out. The results are as follows:

Brand A: $\bar{x}_1 = 37,900$ kilometers; $s_1 = 5100$ kilometers.
Brand B: $\bar{x}_2 = 39,800$ kilometers; $s_2 = 5900$ kilometers.

i) Find the pooled estimate of the population variance.

ii) Using a significance level of $\alpha = 0.05$ and the pooled estimate of variance from part i), test the hypothesis that there is no difference in the average wear of the two brands of tires. (*You are required to write the detail of the test: null and alternative hypotheses, rejection criterion, observed value of the test statistic, p-value, conclusion in plain language - you may use bounds for the p-value.*)

iii) What assumption(s) do you need to make for the above statistical test to be valid?

c) Suppose that X and Y are independent normal variables:

$$X \sim N(-1, 1), \qquad Y \sim N(1, 2)$$

i) What is the distribution of $X + Y$?

ii) Calculate $P(X + Y < 1)$.

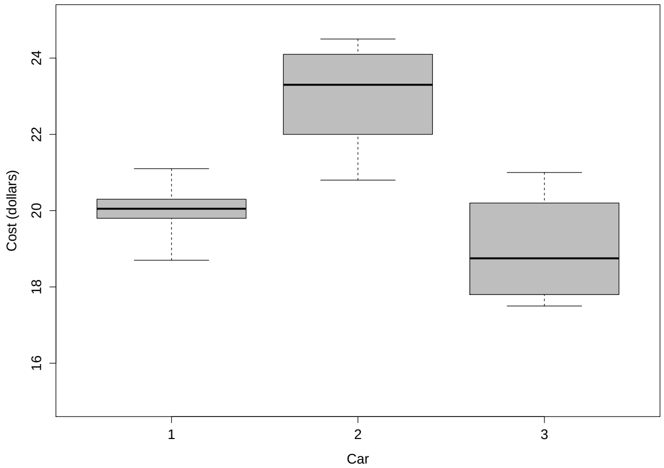
Please see over ...

Answer this question in a separate book marked Question 2

2. [**20 marks**] An experiment is performed to compare the economy of operation of three types of *hybrid* automobiles that operate by both a gasoline engine and electricity. Six autos of each type are driven for 500 miles in the same city, and the variable of analysis is the total cost of gasoline, electricity and maintenance. The results, in dollars, are given in the table below:

Car 1	Car 2	Car 3
20.3	24.5	21.0
19.8	20.8	17.8
21.1	22.0	18.1
18.7	23.1	19.4
20.0	23.5	17.5
20.1	24.1	20.2
$n_1 = 6$	$n_2 = 6$	$n_3 = 6$
$\bar{x}_1 = 20.0$	$\bar{x}_2 = 23.0$	$\bar{x}_3 = 19.0$
$s_1 = 0.780$	$s_2 = 1.383$	$s_3 = 1.421$

Comparative boxplots are given in the figure below.



As part of a cost-benefit analysis, the engineers would like to know if there are any differences in the cost of operating the different types of *hybrid* cars.

a) What do the boxplots tell you about the cost of operating the different types of *hybrid* cars?

b) What assumptions need to be valid for an Analysis of Variance to be an appropriate analysis? Comment on the suitability of these assumptions here, where applicable.

Assume from now on that these assumptions are valid.

Please see over ...

c) An ANOVA table was partially constructed to summarise the data:

Source	df	SS	MS	F
Treatment	**(1)**	26.000	**(2)**	17.18
Error	**(3)**	**(4)**	**(5)**	
Total	**(6)**	27.513		

Copy the ANOVA table in your answer booklet. Complete the table by determining the missing values (1)–(6).

d) Using a significance level of $\alpha = 0.05$, carry out the ANOVA F-test to determine whether the cost of operation is significantly different for the different types of *hybrid* cars. (*You can use the numerical values found in the above table, however you are required to write the detail of the test: null and alternative hypotheses, rejection criterion, observed value of the test statistic, p-value, conclusion in plain language - you may use bounds for the p-value.*)

e) From the previous results, construct a 95% two-sided confidence interval on the difference between the 'true' cost of operating car 1 and car 2, that is, $\mu_1 - \mu_2$. Would you conclude that there is a significant difference between these two means? Explain.

f) The engineers responsible for the study carry out a two-sample t-test to compare the cost of operating car 1 to car 3 and obtain a p-value of 0.17. Does this allow you to come to the same conclusion as the ANOVA F-test in d), at overall level $\alpha = 0.05$? Explain.

Answer this question in a separate book marked Question 3

3. **[20 marks]**

A study of rock fabric, pore geometry and mineralogy (petrological parameters) on transport properties in certain types of rocks includes 24 observations from X, the `permeability` (the ability for gases or fluids to flow through rocks, measured in mD) and Y, capillary `absorption` (measured in $g/(m^2 * \sqrt{s})$), plotted in Figure 1. Source: "Rock fabric, pore geometry and mineralogy effects on water transport in fractured dolostones", *Engineering Geology* 107: 1–15 (2009)

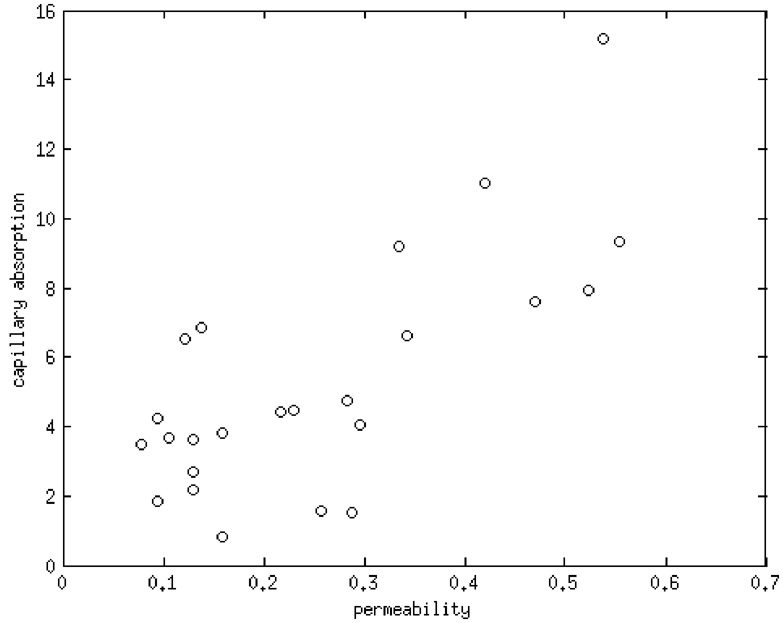


Figure 1: Scatterplot of capillary absorption vs. permeability of 24 rocks.

The researchers fitted a linear regression model:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Using the regression analysis output at the end of the question, answer the following questions.

a) Is the variable `permeability` significant in predicting the variable `absorption`? Test that hypothesis at the 5% significance level. (*You can use the numerical values found in the provided output; however, you are required to write the detail of the test: null and alternative hypotheses, rejection criterion, observed value of the test statistic, p-value and conclusion in plain language.*)

b) Determine the observed sample correlation coefficient between the variables `absorption` and `permeability`.

c) Determine a 95% confidence interval for the change in the mean of `absorption` for an increase of 1 mD in `permeability`.

d) Give a point estimate of the mean of `absorption` when the value of `permeability` is 0.6 mD.

e) Find a 95% confidence interval for the mean value of `absorption` when `permeability` is 0.4 mD.

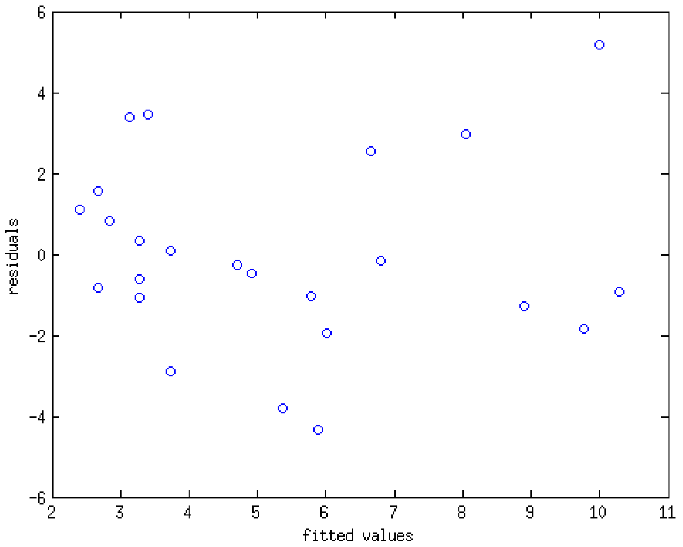
f) Find an interval which you are 95% confident will contain the value of `absorption` when the `permeability` is assumed to be 0.4 mD.

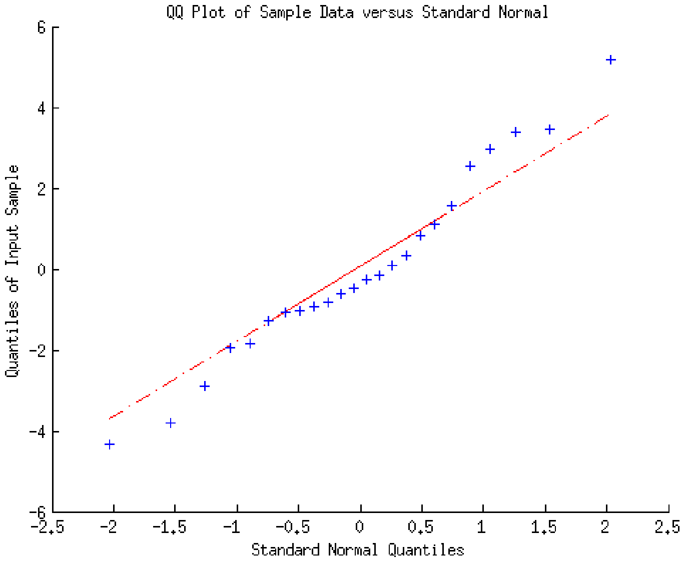
g) Explain why you would expect the interval obtained in Part f) to be different to that in Part e).

h) List three essential assumptions that the error ϵ in the linear regression model must satisfy for the above statistical inferences to be valid. Which of these three can be checked by considering the accompanying regression analysis output? Explain whether these verifiable assumptions are supported.

Regression analysis output for Question 3

```
Regression Analysis: absorption versus permeability

The regression equation is
absorption = 1.1136 + 16.5378 permeability


Predictor         Coef     SE Coef       T          P
Constant        1.1136      0.9592   1.161      0.258
permeability   16.5378      3.2551   5.081   4.34e-05


S = 2.384   R-Sq = 54.0%   R-Sq(adj) = 51.9%


Predicted Values for New Observations

New
Obs       Fit    SE Fit         95% CI              95% PI
  1    7.7287    0.6804  (6.3176, 9.1398)   (2.5865, 12.871)


Values of Predictors for New Observations

New
Obs  permeability
  1  0.4
```


residuals
fitted values
6
4
2
0
-2
-4
-6
2
3
4
5
6
7
8
9
10
11


QQ Plot of Sample Data versus Standard Normal
Quantiles of Input Sample
Standard Normal Quantiles
6
4
2
0
-2
-4
-6
-2.5
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
2.5