RESTRICTED OPEN BOOK EXAMINATION (Not to be removed from the examination hall)
Data provided: "Statistics Tables" by H.R. Neave

MAS5052

The University Of Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS **Spring Semester 2012–2013**

Basic Statistics **2 hours**

RESTRICTED OPEN BOOK EXAMINATION.
Candidates may bring to the examination lecture notes and associated lecture material (including set textbooks) plus a calculator that conforms to University regulations.
Candidates should attempt ***ALL*** *questions.*
The maximum marks for the various parts of the questions are indicated.
The paper will be marked out of 80.

Please leave this exam paper on your desk
Do not remove it from the hall

Registration number from U-Card (9 digits)
to be completed by student

Blank

1 The eruptions of the "Old Faithful" geyser in Yellowstone National Park, Wyoming were recorded between August 1 and August 15, 1985 by Azzalini and Bowman. The lengths of the observed eruptions (in mins) can be found below

Eruption Time (min)	0–1.5	1.5–2	2–2.5	2.5–3.5	3.5–4	4–5	5–5.5
Frequency	1	79	19	12	70	114	4

(i) Represent the data in a suitable graphical format. ***(6 marks)***

(ii) Provide a (very) brief interpretation of the data. ***(2 marks)***

2 Let $\{X_1, \ldots, X_n\}$ and $\{Y_1, \ldots, Y_m\}$ be two independent random samples with $\mathrm{E}[X_i] = \mu$, $\mathrm{Var}[X_i] = \sigma^2$, $\mathrm{E}[Y_j] = 3\mu$, $\mathrm{Var}[Y_j] = \sigma^2/2$.

(i) Prove that

$$T_1(\mathbf{X}, \mathbf{Y}) = \frac{\bar{X} + \bar{Y}}{4}, \quad T_2(\mathbf{X}, \mathbf{Y}) = \frac{3\bar{X} + \bar{Y}}{6} \quad \text{and} \quad T_3(\mathbf{X}, \mathbf{Y}) = \frac{n\bar{X} + 2m\bar{Y}}{n + 6m}$$

are unbiased estimators of μ. ***(6 marks)***

(ii) Explain (with justification) which of these estimators is best if

(a) $n = 2$ and $m = 22$? ***(5 marks)***

(b) $n = 22$ and $m = 2$? ***(5 marks)***

3 The blood pressure of seven randomly chosen individuals was measured in both a supine (lying-down) and a standing position. The results are given below:

Individual	1	2	3	4	5	6	7
Supine Blood Pressure (mmHg)	119	161	126	126	111	135	135
Standing Blood Pressure (mmHg)	120	164	130	128	112	137	138

(i) Is there evidence that the standing blood pressure is different from the supine blood pressure on average? ***(9 marks)***

(ii) State clearly the assumptions underlying the test you have performed and explain how they might be verified [you do not need to perform the checks you suggest]. ***(4 marks)***

(iii) Construct a 95% confidence interval for the average difference in blood pressure measured while supine and standing. ***(3 marks)***

4 Let $X_1, X_2, \ldots, X_n$ be a random sample from a Gaussian distribution with mean zero and variance $1/\lambda$. Show that the most powerful test for

$$H_0 : \lambda = \lambda_0 \quad \text{vs} \quad H_1 : \lambda = \lambda_1 > \lambda_0$$

rejects the null hypothesis if

$$T(\mathbf{X}) = \sum_{i=1}^{n} X_i^2$$

is small enough. ***(8 marks)***

5 A survey of 150 graduates who studied either arts or science subjects at university was conducted to examine their future employment patterns. The data is shown in the following table:

	Faculty		
	Arts	Science	
Public Sector	21	12	33
Private Sector	41	58	99
Unemployed	10	8	18
	72	78	150

How strong is the evidence that the careers of arts and science graduates are different? ***(8 marks)***

6 Suppose that 175 heads and 225 tails resulted from 400 tosses of a coin. Find a 90% confidence interval for the probability of a head. Find a 99% confidence interval. Does this appear to be a fair coin? ***(8 marks)***

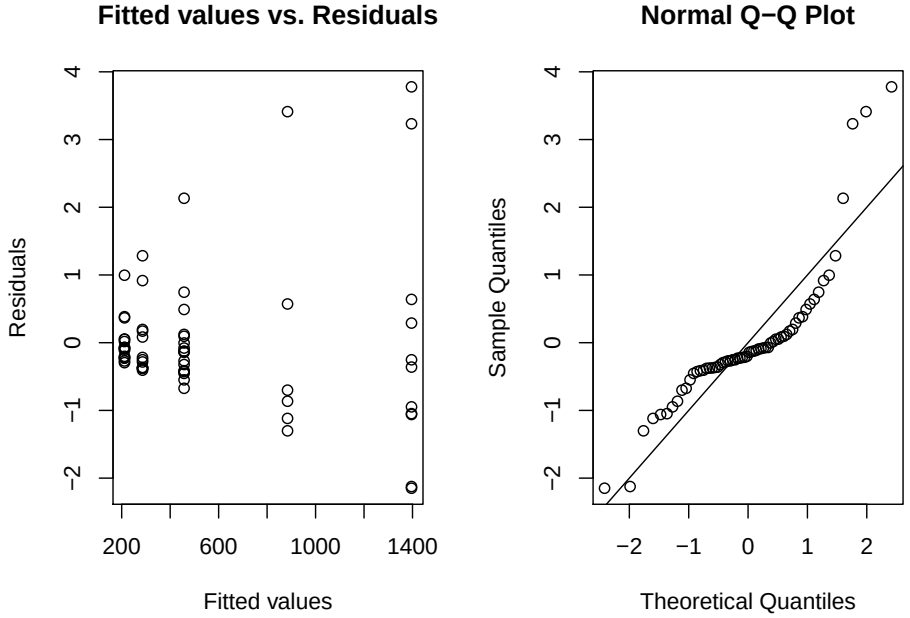
7 The survival time (in days) of 64 patients with advanced cancer of the stomach, bronchus, colon, ovary or breast were measured. The purpose of the study was to determine if patient survival differed with respect to the organ affected by the cancer.

The data were loaded into **R** in the form of a dataframe `cancer` with columns `survival` (the survival time in days) and `organ` (the type of cancer they presented with). The following analysis was run:

```
> lm2 <- lm(formula = survival~organ-1, data = cancer)
> lm1 <- lm(formula = survival~1, data = cancer)
>
> anova(lm1, lm2)
Analysis of Variance Table

Model 1: survival ~ 1
Model 2: survival ~ organ - 1
  Res.Df      RSS Df Sum of Sq      F    Pr(>F)
1     63 37983905
2     59 26448144  4  11535761 6.4334 0.0002295 ***
---
```

Plots were also produced of the residuals from the first model fitted i.e. `lm2`. These are shown below:



(i) Describe the **R** models fitted and interpret the results of the ANOVA.

(3 marks)

(ii) Why in the residual plots are there only 5 different fitted values?

(1 mark)

7 (continued)

(iii) Comment on both of the residual plots. Could you suggest a transformation of the variables which might be better? ***(4 marks)***

8 The yield of a crop (in metric tons), Y, was assumed to be linearly related to the amount of fertiliser used (in Kg), X. The data collected are:

y_i	0.12	0.21	0.34	0.61	0.13	0.17	0.21	0.34	0.62	0.71
x_i	0.7	0.98	1.16	1.75	0.76	0.82	0.95	1.24	1.75	1.95

Using `R` the output of fitting a simple linear regression, $y_i = \alpha + \beta x_i + \varepsilon_i$, is shown below.

```
Residuals:
      Min         1Q     Median         3Q        Max
-0.042951 -0.023396 -0.002618  0.017760  0.052649

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.49708    0.02084   23.86 1.01e-08
 x           2.04889    0.05145   39.83 1.74e-10

Residual standard error: 0.03424 on 8 degrees of freedom
Multiple R-squared: 0.995,Adjusted R-squared: 0.9944
F-statistic:  1586 on 1 and 8 DF,  p-value: 1.738e-10
```

(i) Test the hypothesis $\beta = 1$ vs $\beta \neq 1$, using a test size of 0.05. ***(4 marks)***

(ii) Provide a 99% confidence interval for β. ***(4 marks)***

End of Question Paper