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 2013–2014**

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 A house builder is interested in the income of its property purchasers. It makes three types of property: detached, semi-detached and apartments. It collects data on a sample of its purchasers and summary statistics can be found below:

House Type	Min	Lowest quartile	Median	Upper quartile	Max
Detached	37	49	57	72	95
Semi-Detached	27	41	46	58	73
Apartments	18	22	28	34	51

(i) Provide a suitable display to illustrate the differences in ownership between the three types of property. ***(4 marks)***

(ii) Comment briefly on these differences. ***(4 marks)***

2 Let X be a *single observation* from the density

$$f_X(x;\theta) = \theta\, x^{\theta-1}\ ; \quad 0 < x < 1, \quad \theta > 0\ .$$

The test "Reject H_0 if and only if $X \geq 1/2$" is used to contrast $H_0 : \theta \leq 1$ against $H_1 : \theta > 1$.

(i) Find the size of the test. ***(3 marks)***

(ii) Sketch the power function. ***(5 marks)***

3 Suppose that independent observations X and Y are taken from a $Ga(a, 1/\eta)$ and $Ga(b, 1/\eta)$ distribution respectively where both a and b are known and positive.

Note: The density of a $Ga(\alpha, \beta)$ random variable is

$$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

(i) Find the maximum likelihood estimate of η ***(5 marks)***

(ii) Show that the maximum likelihood estimator of η is unbiased and find its variance ***(4 marks)***

(iii) Compare the mle estimator with the alternative estimator

$$T_1 = \frac{1}{2}\left(\frac{X}{a} + \frac{Y}{b}\right).$$

Which is better? ***(7 marks)***

4 There are good reasons for thinking that the time required to recover a normal pulse rate after physical activity maybe longer in the evening than in the morning. Ten people of similar age, sex and size were arbitrarily selected, asked to run at a moderate pace for 30mins and then the time each took to recover a normal pulse rate was measured on both morning and evening. The results (in mins) were as follows:

Person no.	1	2	3	4	5	6	7	8	9	10
Evening	5.3	7.1	6.4	6.6	4.7	3.6	6.5	4.9	3.4	4.0
Morning	4.5	6.9	6.5	4.7	3.2	4.0	5.0	5.4	2.0	4.1

The R output from two analyses of these data is presented below:

Analysis 1:

```
t.test(x=evening,y=morning,var.equal=T,paired=F)

Two Sample t-test

data:  evening and morning
t = 0.9899, df = 18, p-value = 0.3353
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.6958699  1.9358699
sample estimates:
mean of x mean of y
     5.25      4.63
```

Analysis 2:

```
t.test(x=evening,y=morning,var.equal=T,paired=T)

Paired t-test

data:  evening and morning
t = 2.1716, df = 9, p-value = 0.05796
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.02584928  1.26584928
sample estimates:
mean of the differences
                   0.62
```

(i) Which of these analyses is more appropriate? Give reasons for your answer. ***(2 marks)***

(ii) What other checks would you need to carry out before deciding your preferred analysis is applicable? ***(3 marks)***

(iii) Assuming these checks prove satisfactory, do the data provide evidence that time to recovery is larger in the evening that in the morning? ***(3 marks)***

5 You are playing a game with your friend which involves tossing four coins and counting the total number of heads. Your friend has brought the coins himself and you don't entirely trust him. You play the game 100 times, with the following observed counts:

Number of heads	Number of occurences
0	3
1	24
2	36
3	29
4	8

Perform a test of whether the coins are fair. What do you conclude? ***(8 marks)***

6 A Human Resources Manager at a UK company with 38 male and 49 female employees decides to commission a study to learn about employee satisfaction with the workplace facilities. Assume that a decision has already been made to sample 10% of the employees and to stratify the study by gender.

(i) Why might the organisers of the study have decided to stratify by gender? ***(2 marks)***

(ii) Suppose that there is a particular concern about the satisfaction of female employees. If the sampling fraction for female employees is to be raised to 15%, but the overall sampling fraction is to remain the same, how many male and female employees will be sampled? ***(3 marks)***

(iii) One of the questions that the HR Manager proposes to ask is "How much did you spend in the on-site catering facility last week?". Comment on this choice of question. ***(3 marks)***

7 Let $X_1, \ldots, X_n$ be independent observations. Suppose X_i has a $Po(\mu_i)$ distribution for $i = 1, \ldots, n$.

(i) Show that, according to the Neyman-Pearson Lemma that the most powerful test for

$$H_0 : \mu_i = 1 \quad 1 \leq i \leq n$$

against

$$H_1 : \mu_i = \begin{cases} 1 & 1 \leq i \leq r \\ 2 & r < i \leq n \end{cases}$$

for known r with $1 < r < n$ rejects the null hypothesis if

$$T(\mathbf{X}) = \sum_{i=r+1}^{n} X_i \geq k$$

for some k ***(6 marks)***

(ii) What is the distribution of $T(\mathbf{X})$ under H_0? ***(2 marks)***

8 In a car parts factory, the operation temperature of one of the machines must be controlled to keep the production line working. When a certain critical temperature is reached, a continuous flow of coolant is passed through the machine until the optimal working temperature is re-established. The operator gathered data on 53 randomly chosen days comprising the log-amount of coolant used (`Logcool`) and the time taken to recover optimal temperature (`Recoverytime`). The following output was obtained (from `R`).

```
lm(formula = Recoverytime ~ Logcool)

Residuals:
    Min      1Q  Median      3Q     Max
-25.390 -10.332  -2.223   6.784  43.572

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -9.233     12.765  -0.723   0.4728
Logcool       16.026      6.317   2.537   0.0143 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 15.5 on 51 degrees of freedom
Multiple R-squared: 0.1121,     Adjusted R-squared: 0.09466
F-statistic: 6.437 on 1 and 51 DF,  p-value: 0.01428
```

(i) Explain carefully what model has been fitted and what assumptions have been made. Give the estimates of the model parameters. ***(8 marks)***

(ii) The operator wants to know whether the amount of coolant gives any indication of the length of the recovery time. Assuming the model is acceptable, give a conclusion on this that is appropriate for the operator. ***(4 marks)***

(iii) Use the model to give an estimate of the mean recovery time when the log amount is 4.00. What additional information do you need in addition to the output above in order to calculate a standard error for this estimate? ***(2 marks)***

(iv) Do any of the diagnostic results from the analysis give any grounds for concern about the model assumptions? Explain your answer. ***(2 marks)***

End of Question Paper