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 2010–2011**

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 Data arising from a study into fuel consumption in miles per gallon of a sample of 85 cars are recorded in the following frequency table.

Consumption (mpg)	10–20	20–30	30–35	35–40	40–50
Frequency	2	25	24	25	9

(i) Represent the data in a histogram. ***(6 marks)***

(ii) Provide a (very) brief interpretation of the data based on your graph. ***(2 marks)***

2 The vice-chencellor at Sheffield University wants to commission a study to investigate the impact of proposed tuition fee increases on new 1^{st} year Mathematics students. Of the new entrants, 80 obtained an A* in their A-level Mathematics (the top mark available); 120 obtained an A; and 40 obtained a B. After consultation, a decision was made to sample 15% of the students and to stratify by their A-level mark.

(i) Why might the vice-chancellor have decided on stratification by A-level mark? ***(2 marks)***

(ii) Suggest a suitable proportional sampling scheme for the study outlined above. ***(2 marks)***

(iii) The vice-chancellor is particularly interested in the opinions of those students who obtained either an A* or a B in their A-levels. He wishes to raise the sampling fraction for both these sets of students to 20%, but keep the overall sampling fraction the same at 15%. How many students in each group will now be sampled? ***(2 marks)***

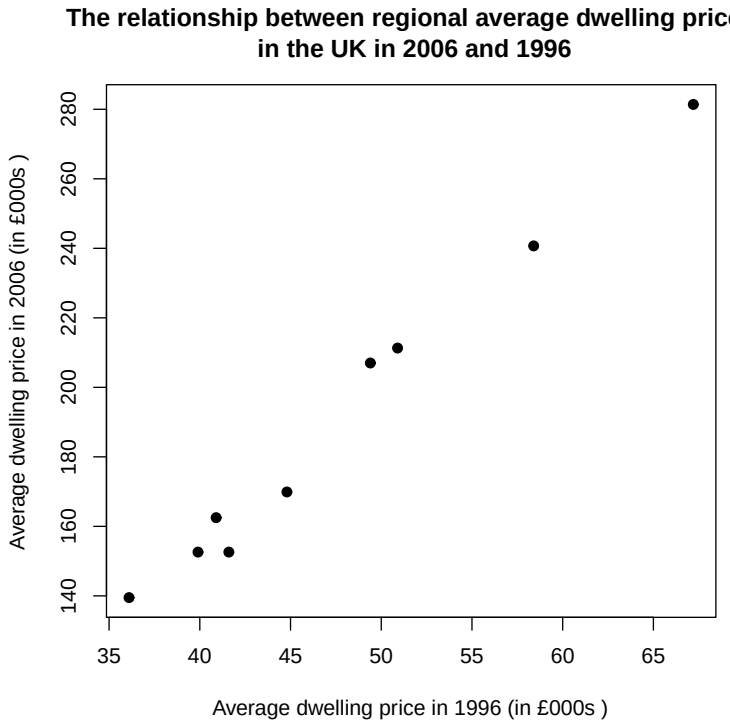
(iv) The vice-chancellor wishes to include in the survey the following question,

"Without increased funding of some form, the university faces a large budget deficit which must be tackled. As students benefit from the education they receive, do you agree that asking them for an increased financial contribution is reasonable?"

Comment on this choice of question. ***(2 marks)***

3 Data from Social Trends 28 (1998) and Social Trends 38 (2008), relating to average dwelling prices (in thousands of £s) in regions of England in 1996 and 2006, can by summarised as follows.

Region of England	1996 x	2006 y
North East	36.1	139.5
North West	41.6	152.6
Yorkshire and the Humber	39.9	152.6
East Midlands	40.9	162.5
West Midlands	44.8	169.9
East	50.9	211.3
London	67.2	281.4
South East	58.4	240.7
South West	49.4	207.0



Summary statistics are:

$$\sum x = 429.2, \sum y = 1717.5,$$
$$\sum x^2 = 21263.2, \sum y^2 = 345925.2, \sum xy = 85678.55.$$

(i) Describe the relationship between house prices in 2006 and 1996 and explain why the 2006 price is shown on the vertical axis. ***(3 marks)***

(ii) Find the least squares estimate of the best-fit straight line for these data. ***(5 marks)***

3 (continued)

(iii) Test the hypothesis that there is no relationship between house prices in 1996 and 2006. ***(6 marks)***

[Hint: For a simple linear model (i.e. $y_i = \alpha + \beta x_i + \epsilon_i$) we have $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = s_{yy} - \hat{\beta}^2 s_{xx}$.]

(iv) What check would you carry out on the data to check that your linear model is appropriate and how would you do this? ***(2 marks)***

4 A pack of cards contains 52 cards split evenly between 4 different suits — hearts, spades, diamonds and clubs. A psychic claims to be able to foresee the suit of a card before it is drawn. 120 such cards are drawn at random (with replacement) from a pack, with the psychic giving a prediction on each occasion. Let X be the number of correct predictions.

(i) Suggest a statistical model for X, and a null hypothesis that could be tested regarding the psychic's claim. ***(2 marks)***

(ii) The psychic correctly predicts the suit of the card 37 times out of 120. He states:

"If I were guessing, I should only get 30 predictions right, and the probability of getting 37 right would be only 0.028. This is a really small probability (smaller than 0.05) so I am clearly not guessing"

What is wrong with this argument? ***(2 marks)***

(iii) Given the observation of $X = 37$ conduct a two-sided hypothesis test of your hypothesis in part (i). ***(4 marks)***

5 Gregor Mendel was one of the first people to investigate inheritance of genetic traits by crossing pea plants, specifcally looking at two traits — shape and colour. According to his theory, the traits should be present in the following proportions

- Round and Yellow: 9/16
- Round and green: 3/16
- Angular and Yellow: 3/16
- Angular and green: 1/16

After collecting data from 556 plants, he observed the following

Shape	Colour	Number Observed
Round	Yellow	315
Round	Green	108
Angular	Yellow	101
Angular	Green	32

Carry out a suitable test of his genetic theory. ***(8 marks)***

6 Suppose that $X_1, X_2, \ldots, X_n$ are independent random variables, each following a $Unif[0, \theta]$ distribution.

(i) Show the maximum likelihood estimator is $\hat{\theta} = \max_{i=1,\ldots,n} X_i$. ***(5 marks)***

(ii) Calculate $P(\hat{\theta} \leq y)$. Hence show that $E[\hat{\theta}] = \dfrac{n}{n+1}\theta$. ***(7 marks)***

[Hint: If $\max_{i=1,\ldots,n} X_i \leq y$, what can we say about the value of each X_i?]

(iii) Consider an alternative estimator $\bar{\theta} = 2\bar{X}$. Is this an unbiased estimator of θ? Is it a reasonable estimator? ***(4 marks)***

7 A survey has been undertaken to investigate methods of rice growing. Three different types of rice have been used together with three different cultivation techniques. Each combination of rice and cultivation technique has been grown three times in identical fields. We record the total amount of rice (in kg) produced in each field.

	Type 1	Type 2	Type 3
Technique A	9.8, 10.1, 9.8	9.2, 8.6, 9.2	8.4, 7.9, 8.0
Technique B	9.9, 9.5, 10.0	9.1, 9.1, 9.4	8.6, 8.0, 8.0
Technique C	11.3, 10.7, 10.7	10.3, 10.7, 10.2	9.8, 10.1, 10.1

Interpret the following R output, making sure you give the model initially fitted and the most appropriate model for the data. Explain your reasoning.

```
fit1 <- aov(formula = yield ~ technique * type, data = rice))
summary(fit1)
                Df  Sum Sq Mean Sq F value    Pr(>F)
technique        2 11.7800  5.8900 80.3182 1.071e-09 ***
type             2  9.2600  4.6300 63.1364 7.325e-09 ***
technique:type   4  0.7400  0.1850  2.5227   0.07709 .
Residuals       18  1.3200  0.0733
```

(8 marks)

8 Let $X_1, X_2, \ldots, X_n$ be a random sample from a Bernoulli distribution with probability of success p. We wish to test

$$\begin{aligned} H_0 : p &= p_0 \\ H_1 : p &= p_1 \end{aligned}$$

where $p_0 < p_1$. Show that the most powerful test is to reject H_0 if

$$T = X_1 + X_2 + \ldots + X_n \geq k^\star$$

for some $k^\star$. ***(8 marks)***

End of Question Paper