Data provided: Neave's Tables

MAS370

The University Of Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS **Spring Semester 2011–2012**

MAS370 Sampling Theory and Design of Experiments **2 hours**

Restricted Open Book Examination.
Candidates may bring to the examination lecture notes and associated lecture material (but no textbooks) plus a calculator which conforms to University regulations.
*Marks will be awarded for your best **three** answers. Total marks 90.*

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 An investigator is studying the dependence of a variable Y on two continuous explanatory variables x_1 and x_2, which have been scaled to lie between -1 and 1. It is known that $EY = 0$ when both $x_1 = 0$ and $x_2 = 0$, and the following model is proposed. Each observation is subject to a measurement error with mean 0 and variance σ^2.

$$EY = \beta_1 x_1 + \beta_2 x_2.$$

The investigator proposes to take four observations, at (-1,0), (1,0), (0,-1) and (0,1). Denote the four observations by $Y_1, \ldots, Y_4$.

(i) Find the least squares estimators of β_1 and β_2, verify that they are unbiased, and give their variances in terms of σ^2 only. ***(9 marks)***

(ii) By examining the form of your estimators in part (i) (rather than by considering the design matrix), briefly explain why β_1 and β_2 are orthogonal to each other for the chosen design. ***(4 marks)***

(iii) Show that this design is neither D-optimal nor G-optimal, by using the General Equivalence Theorem. ***(7 marks)***

(iv) Suggest an alternative design, with four observations, that is D-optimal. Justify your suggestion. ***(10 marks)***

2 An experiment is to be carried out to investigate the effect of three diets and three drugs on blood pressure. Nine volunteers are recruited to the study, and are grouped by weight into three blocks of three. A design is chosen based on the following Latin square.

A	B	C
C	A	B
B	C	A

The following model is to be fitted to the data.

$$EY_{ij} = \mu + \theta_i + \phi_j + \psi_{\kappa(i,j)},$$

with $i = 1, 2, 3$ representing block, $j = 1, 2, 3$ representing diet, and $\psi_{\kappa(i,j)}$ representing the effect of the drug given to the individual on diet j in block i (so that $\kappa(i, j) = 1, 2$ or 3). The constraints

$$\sum_{i=1}^{3} \theta_i = \sum_{j=1}^{3} \phi_j = \sum_{k=1}^{3} \psi_k = 0,$$

are applied.

(i) Write this model in matrix notation, and explain, with justification, which groups of parameters are orthogonal to each other. ***(12 marks)***

(ii) Suppose the experiment is to be extended to investigate the effect of three different exercise regimes, in addition to drug, diet and weight. State how you would allocate exercise regimes to volunteers using a second Latin square that is orthogonal to the first. State how you would modify the original model, and give the extra columns of the design matrix.
(13 marks)

(iii) In an alternative experiment, the effect of the three diets are to be considered only, but volunteers will still be blocked by weight. If blocks of size two are to be used, how many volunteers are required for the smallest possible balanced incomplete block design? For the smallest possible such design, list which diets should be used in each block. ***(5 marks)***

3 An experiment is to be conducted to investigate the effect of four continuous factors, represented by x_1, x_2, x_3 and x_4, on response variable Y.

(i) Construct a Box-Behnken design with one centre point. ***(8 marks)***

(ii) Now suppose each factor is restricted to one of two levels.

(a) If a complete factorial design is to be used, how many observations would there be? Write down the fullest possible linear model that could be fitted to the data. ***(4 marks)***

(b) Now suppose blocking is to be used, with four blocks, in conjunction with a complete factorial design. If the two block generators are x_1x_2 and x_3x_4, write down the design used within one of the four blocks. Which effects will be confounded with the block effects?

(5 marks)

(c) Give a fractional factorial design for this experiment using the design generator $x_3x_4 = 1$. Find the alias structure, and suggest the most appropriate model to fit. State the resolution of the design. Give one criticism of this design generator, and suggest a better choice.

(13 marks)

4 (i) A small survey has been conducted to estimate the proportion of the population in favour of raising the female retirement age. The sex of each participant in the survey was recorded, and the results are given below.

	males	females
in favour	23	15
against	18	44

If the sample was drawn using simple random sampling, suggest two different estimates of the population proportion in favour of raising the female retirement age. Briefly explain your reasoning. ***(5 marks)***

(ii) An opinion poll is to be taken to estimate the proportion of the adult population in Scotland who are in favour of Scotland leaving the United Kingdom. If a simple random sample is to be used, how large would the sample need to be to ensure that a 90% confidence interval for the true proportion was no wider than 0.1? You may ignore the finite population correction. ***(7 marks)***

(iii) 100 motors have various different ages, all of which are known. It is believed that the remaining lifetime of each motor will be dependent on the motor's age. Five motors are selected at random, their ages are noted and their remaining lifetimes are observed by running them continuously until failure. Let Y_i be the age of the i-th motor, and X_i be the remaining lifetime of the i-th motor. The following summary statistics are observed.

$$\sum_{i=1}^{100} Y_i = 9024, \qquad \sum_{i=1}^{5} y_i = 350, \qquad \sum_{i=1}^{5} x_i = 4646,$$

$$\sum_{i=1}^{5} (x_i - \bar{x})(y_i - \bar{y}) = -13754, \qquad \sum_{i=1}^{5} (y_i - \bar{y})^2 = 13960,$$

where $(x_1, y_1), \ldots, (x_5, y_5)$ are the observed age and remaining lifetime for 5 randomly selected motors.

Suggest two possible estimates of $\bar{X}$, explaining your reasoning. Without referring explicitly to the formulae of your estimators, give an intuitive explanation for the discrepancy between the two estimates. ***(9 marks)***

4 (continued)

(iv) A survey is to be taken to estimate the mean starting wage of individuals following completion of a new training course. Two pilot studies have been conducted. In the first study, a simple random sample of size 20 was used, and the following summary statistics were observed.

$$\sum_{i=1}^{20} x_i = 564.9, \qquad \sum_{i=1}^{20} x_i^2 = 16131.2,$$

where each x_i is measured in £1000.

In the second study, cluster sampling was used. Courses are run at different training centres around the country, with 10 students at each centre. Two training centres were selected as the clusters. The following summary statistics were observed.

$$\sum_{j=1}^{10} x_{1j} = 258.7, \qquad \sum_{j=1}^{10} x_{1j}^2 = 6714.3,$$
$$\sum_{j=1}^{10} x_{2j} = 305.0, \qquad \sum_{j=1}^{10} x_{2j}^2 = 9319.2,$$

where x_{ij} is observation j within cluster i.

Based on the pilot survey data, would you recommend the use of cluster sampling or simple random sampling for the new survey? Briefly explain your reasoning. ***(9 marks)***

End of Question Paper