The University Of Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS | **Spring Semester 2013–2014**

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.
*Answer **all** questions. Total marks 60.*

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 small experiment is being conducted to compare two new treatments against a placebo. There are six participants in the study. Two participants are given the placebo, two are given treatment A, and two are given treatment B. Each observation is subject to a measurement error with mean 0 and variance σ^2. The following model is proposed.

$$EY_{ij} = \mu + \tau_i,$$

for $i = 1, 2, 3$, $j = 1, \ldots, n_i$, with $n_1 = n_2 = n_3 = 2$. The constraint $\tau_1 + \tau_2 + \tau_3 = 0$ is applied.

(i) Write down the design matrix for this design. ***(3 marks)***

(ii) Describe which parameters or groups of parameters are orthogonal. ***(2 marks)***

(iii) Suppose instead that four participants are given the placebo, one is given treatment A and one participant is given treatment B. Describe which parameters or groups of parameters are orthogonal for this design. ***(2 marks)***

(iv) Suppose instead that there are 10 participants in total of which $10 - 2t$ participants are given the placebo, t are given treatment A and t are given treatment B. The following model is proposed

$$EY_{ij} = \tau_i,$$

for $i = 1, 2, 3$, $j = 1, \ldots, n_i$, with $n_1 = 10 - 2t$, $n_2 = n_3 = t$.

(a) Explain why there is no need to impose constraints in this model. ***(1 mark)***

(b) Show that

$$(X^T X)^{-1} = \begin{pmatrix} 1/(10-2t) & 0 & 0 \\ 0 & 1/t & 0 \\ 0 & 0 & 1/t \end{pmatrix}.$$

Find the value of $t \in \{1, 2, 3, 4\}$ that minimizes $Var(\hat{\tau}_2 - \hat{\tau}_1)$ and give the minimum value of the variance. ***(7 marks)***

1 (continued)

(v) Suppose we still want to compare the effect of placebo and the two treatments (labelled P, A, B respectively) but there are now nine participants. In addition there are also three different experimental conditions (labelled $1, 2, 3$). The nine participants are to be allocated to blocks based on whether they are a current smoker, ex-smoker or have never smoked (labelled $1, 2, 3$ respectively). Give 2 orthogonal Latin square designs for this set up. State your design by completing the following table for each design. ***(4 marks)***

participant	smoking status $(1, 2, 3)$	experimental condition $(1, 2, 3)$	treatment (P, A, B)
1			
$\vdots$			
9			

(vi) Does a third orthogonal Latin square exist? Justify your answer. ***(1 mark)***

2 (i) An experiment is to be carried out to investigate the effect of four teaching methods on chemistry exam scores. There are 12 randomly selected students in the study, who will each be taught using one of the four methods. After the course finishes, each participant will be given a chemistry test, and their exam scores will be recorded. The experimenter decides to organise the students into blocks, according to their abilities.

(a) If the four teaching methods are labelled A, B, C, D, explain why the following design satisfies the requirements of a balanced incomplete block design with block sizes of 3.

Block 1: ABC
Block 2: ABD
Block 3: BCD
Block 4: ACD

(1 mark)

(b) For the design in (a), write out the observation vector, parameter vector and design matrix in full and specify any parameter constraints. ***(5 marks)***

(c) Suppose instead that the experimenter is only interested in the proportion of students who would get a mark of 60% with each method. The experimenter proposes to randomly allocate students to methods (with equal numbers of students per method), and observe the sample proportions of students getting 60% or higher for each of the four methods. How many students would be needed **in total**, such that the width of a 90% confidence interval for any single proportion was no wider than 0.2? You may ignore the finite population correction. Part of the following R output will help you.

```
> qnorm(c(0.9,0.95,0.975),0,1)
[1] 1.281552 1.644854 1.959964
```

(5 marks)

 Question 2 continued on next page

2 (continued)

(ii) 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. Each observation is subject to a measurement error with mean 0 and variance σ^2. The following model is proposed.

$$EY = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

The investigator proposes to take four observations of (x_1, x_2) given by $(-1, 1), (1, 1), (0, 0)$ and $(0, -1)$. Denote the response for the four observations as Y_1, Y_2, Y_3, Y_4 respectively.

(a) Show that this design is neither D-optimal nor G-optimal, by using the General Equivalence Theorem. You may use the fact that

$$\begin{pmatrix} 4 & 0 & 1 \\ 0 & 2 & 0 \\ 1 & 0 & 3 \end{pmatrix}^{-1} = \frac{1}{22}\begin{pmatrix} 6 & 0 & -2 \\ 0 & 11 & 0 \\ -2 & 0 & 8 \end{pmatrix}.$$

(6 marks)

(b) Suggest an alternative design with four observations of (x_1, x_2) that is both D-optimal and G-optimal. ***(3 marks)***

3 (i) An ecologist wishes to estimate the population of rainbow trout in a lake. 50 trout are caught, tagged, and released to mix within the population. Another 100 trout are then caught, and 4 are observed to have tags. Estimate the population of trout, and give a standard error for your estimate. State one assumption that you have made when calculating your estimate.

(3 marks)

(ii) A survey is conducted to estimate the mean income of graduates from a university, five years after graduation.

(a) If stratified sampling is to be used, suggest a suitable choice of strata

(1 mark)

(b) A previous survey produced the following data

Stratum	Population size	Sample size	std. dev. (£)	mean (£)
1	5000	50	2000	30000
2	2000	50	500	22000
3	2000	20	1000	25000
4	1000	10	5000	40000

Estimate the mean income for all graduates. ***(2 marks)***

(c) If a new survey is to be conducted with a sample size of 500, suggest a sample size for each stratum using each of the following methods:

- proportional allocation; ***(2 marks)***
- Neyman allocation; ***(2 marks)***
- minimising the variance of the stratified sample mean, subject to unequal costs: a graduate from stratum 4 is four times as expensive to sample as a graduate from a different stratum; costs for the other three strata are the same. ***(3 marks)***

(You may round so that the total sample size is either 499 or 501, if necessary).

(iii) A computer model is given by the function

$$Y = X_1 + X_2X_3.$$

The true values of the three inputs are uncertain, with $X_1 \sim N(1, 2)$, $X_2 \sim N(0, 6)$ and $X_3 \sim N(2, 1)$. The model user can choose to learn the true value of one of the inputs. The model user wishes to reduce the variance of the output Y as much as possible.

(a) Calculate the main effect indices of each input, and hence state which is the best input to learn. You may leave your answers in terms of $Var(Y)$.

(b) Suppose, instead, the model user can choose to learn the value of two of the inputs. Is it best to choose the two with the largest main effect indices from part (a)? Justify your answer.

(7 marks)

End of Question Paper