The University Of Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS **Spring Semester 2014–2015**

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 An investigator is studying the dependence of a variable Y on one continuous explanatory variable x, which has been scaled to lie between -1 and 1. The following model (called model 1) is proposed:

$$EY = \beta_1 x + \beta_2 x^4.$$

The investigator proposes to take five observations, at $x = -1, -1/2, 0, 1/2, 1$.

(i) Show that β_1 and β_2 are orthogonal to each other in model 1. ***(2 marks)***

(ii) Give the variances of the least squares estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ in terms of σ^2, the error variance of each observation. ***(3 marks)***

(iii) Suppose model 2 is formed by adding a parameter β_0 to model 1 to represent the intercept. A statistician claims that the estimates of β_1 would be identical in models 1 and 2 because β_1 and β_2 are orthogonal. Justify whether the statistician's reasoning is correct and whether the estimates of β_1 would be identical in model 1 and model 2. ***(3 marks)***

(iv) By using the General Equivalence Theorem, or otherwise, show that the design $x = -1, -1/2, 0, 1/2, 1$ for model 1 is neither D-optimal nor G-optimal. ***(5 marks)***

(v) Show that no D-optimal or G-optimal design of the form $x = -m, -m, 0, m, m$, exists for model 1. ***(7 marks)***

2 In this question, balanced incomplete block design is abbreviated to BIBD. Using the same definitions as in the course notes, we define the following variables in a BIBD.

$$
\begin{aligned}
t &= \text{number of treatments} \\
k &= \text{number of units in a block} \\
b &= \text{number of blocks} \\
r &= \text{number of applications of each treatment} \\
\lambda &= \text{number of times each pair of treatments appears together in a block}
\end{aligned}
$$

(i) For the unreduced design, justify why $r = \begin{pmatrix} t-1 \\ k-1 \end{pmatrix}$ and $\lambda = \begin{pmatrix} t-2 \\ k-2 \end{pmatrix}$. ***(2 marks)***

(ii) Find the smallest number of blocks for a BIBD with $t = 5$ and $k = 2$. ***(3 marks)***

(iii) Write down the BIBD in part (ii) if the 5 treatments are labelled A, B, C, D and E. ***(2 marks)***

(iv) Give a BIBD with $t = 5$ and $k = 4$ using an appropriate Latin Square or otherwise. Verify that this design satisfies the requirements of a BIBD. ***(4 marks)***

(v) Consider a BIBD in which $Y_{ij} = \beta_i + \tau_j + \epsilon_{ij}$, where i is the block and j is the treatment and $\epsilon_{ij} \sim N(0, \sigma_1^2)$. Consider also a completely randomised design in which $Y_{jk} = \tau_j + \epsilon_{jk}$, where Y_{jk} is the k-th observation in the j-th treatment group and $\epsilon_{jk} \sim N(0, \sigma_2^2)$. Calculate the ratio of $\mathrm{var}\,(\hat{\tau}_j)$ in the BIBD in part (iv) to $\mathrm{var}\,(\hat{\tau}_j)$ in a completely randomised design with the same number of observations (putting the variance for the BIBD estimator in the numerator). Would you expect this ratio to be less than or more than 1? Justify your answer. ***(4 marks)***

(vi) A computer model is given by the function

$$Y = X_1^2 + 2X_2X_3.$$

The true values of the three inputs are uncertain, with $X_1 \sim N(1, 1)$, $X_2 \sim U[0, 1]$ and $X_3 \sim N(2, 4)$, with X_1, X_2, X_3 independent. 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. Suggest which input the model user should choose to learn, justifying your reasoning. You may use the results that $E(X_1^4) = 10$, and for $X \sim U[a, b]$, $Var(X) = (b - a)^2/12$. ***(5 marks)***

3 (i) An opinion poll has been conducted to estimate the proportion of the UK population (aged 18 or over) in favour of leaving the European Union (EU). A stratified sample has been taken, with four strata.

Stratum	Population size (millions)	Sample size	Number in favour of leaving EU
1	40	200	110
2	5	200	65
3	3	200	72
4	2	200	95

(a) Estimate the proportion of the UK population (aged 18 or over) in favour of leaving the European Union. ***(1 mark)***

(b) A new poll is to be taken, again using stratified sampling, but this time using proportional allocation. The standard deviation of the estimated proportion is required to be no larger than 0.01. How large should the total sample size be? Ignore the finite population correction. ***(6 marks)***

(ii) A new test has been devised to assess reading comprehension in Year 3 school pupils (aged 7-8). In the population of interest there are 1000 schools. Each school has 50 Year 3 pupils. A random sample of 20 schools is selected to try the test. Within each of the 20 selected schools, all 50 Year 3 pupils take the test. Let x_{ij} be the score (out of 100) of the j-th pupil within the i-th selected school in the sample. If it is given that

$$\sum_{i=1}^{20}\sum_{j=1}^{50} x_{ij} = 56260,$$

$$\sum_{i=1}^{20} \bar{x}_i^2 = 64893,$$

estimate what the mean score would be for the population of pupils, if all 50000 Year 3 pupils were to take the test. Calculate an estimated standard error for your estimate. ***(6 marks)***

 Question 3 continued on next page

3 (continued)

(iii) An ecologist wishes to estimate the total number of females of a particular species across 10 distinct regions. In region i, the total number of animals y_i in the species is counted, for $i = 1, \ldots, 10$. The total number of females x_i is counted for regions $i = 1, \ldots, 5$. The following data are obtained.

i	1	2	3	4	5	6	7	8	9	10
y_i	75	80	100	50	90	20	10	80	65	40
x_i	30	30	35	20	50	-	-	-	-	-

(a) The ecologist proposes to double the observed number of females, to get the estimated total for all 10 regions. Give one criticism of this suggestion. ***(1 mark)***

(b) Calculate an alternative estimate that you believe to be more appropriate, justifying your reasoning. Without doing any further calculation, give a formula you would use to calculate an estimated standard error for your estimate, defining any notation you introduce carefully. ***(6 marks)***

End of Question Paper