Data provided: Neave's Tables

The University Of Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS Spring Semester 2010–2011

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 one continuous explanatory variable x_1, which has been scaled to lie between -1 and 1. It is known that $EY = 0$ when $x_1 = 0$, and the following model is proposed.

$$EY = \beta_1 x_1 + \beta_{11} x_1^2.$$

The investigator proposes to take four observations, at $x_1 = -1, -0.5, 0.5, 1$.

(i) Show that β_1 and β_{11} are orthogonal to each other. ***(4 marks)***

(ii) If each observation is subject to a measurement error with mean 0 and variance σ^2, give the variances of the least squares estimators $\hat{\beta}_1$ and $\hat{\beta}_{11}$ in terms of σ^2. ***(2 marks)***

(iii) If a constant β_0 were to be added to the model, explain why β_0 would not be orthogonal to β_{11} for any choice of design points (with at least one design point not at 0). ***(3 marks)***

(iv) Show that this design (for the model $EY = \beta_1 x_1 + \beta_{11} x_1^2$) is neither D-optimal nor G-optimal, by using the General Equivalence Theorem. ***(5 marks)***

(v) Suggest an alternative design, with four observations, that is D-optimal. Justify your suggestion. ***(9 marks)***

(vi) The investigator is interested in predicting Y at values of x close to its upper limit, and wishes to consider V-optimality with the weight function $w(x) = 1 + x$. Show that your design in part (v) is V-optimal out of all possible orthogonal designs with 4 observations. ***(7 marks)***

2 (i) In a small study, four treatments A, B, C and D are to be compared. There are 12 females in the study, all of different ages. It is suspected that age may have an effect on the response variable Y. Each person in the study can only receive one treatment.

(a) Explain how a balanced incomplete block design with 6 blocks could be used in the design of the study. How would you choose the blocks, and which treatments would you used in each block? What would you randomise in your design? ***(9 marks)***

(b) The following model is proposed for the data.

$$Y_{it} = \gamma_i + \delta_t + \varepsilon_{it},$$

with Y_{it} the observed response variable on treatment t in block i, with $t = A, B, C$ or D, and $\varepsilon_{it} \sim N(0, \sigma^2)$. If $\delta_A + \delta_B + \delta_C + \delta_D = 0$, give an expression for the estimator of δ_C in terms of Y_{it}, and state the variance of the estimator. ***(3 marks)***

(c) Suppose instead there were 20 females. Why would a balanced incomplete block design not be possible? ***(4 marks)***

(d) As an alternative to blocking, it is suggested to allocate each treatment to three patients at random, and include terms to represent the effect of age in the linear model. Give one disadvantage of this approach. ***(3 marks)***

(ii) An alloy is being developed from a mixture of four different metals A, B, C and D. A single observation consists of choosing proportions of each metal, for example, 50% metal A, 20% metal B, 20% metal C and 10% metal D, constructing the alloy, and then measuring how much weight the alloy can bear before breaking.

For a design with 10 observations, state the most complex linear model that would be appropriate to fit to the data, justifying your answer, and suggest a suitable design. ***(11 marks)***

3 (i) Construct a central composite design for two factors x_1 and x_2 and 12 observations, that would be suitable for fitting the model

$$EY = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{111} x_1^3 + \beta_{222} x_2^3.$$

Briefly justify your choice of design points.

(6 marks)

(ii) An experiment is to be conducted to investigate the effect of four factors, represented by x_1, x_2, x_3 and x_4 on response variable Y. Each factor can take one of two levels, with the two levels denoted by -1 and $+1$.

(a) Give a fractional factorial design for this experiment using the design generator $x_1 x_2 x_4 = 1$. Find the alias structure, and suggest the most appropriate model to fit. State the resolution of the design.
(13 marks)

(b) Suppose instead that the first factor x_1 is qualitative and takes one of four levels. If the experimenter does not wish to consider any interaction terms, suggest a design with 8 observations that will enable the estimation of the effects of all four factors, with orthogonality between factors. Hint: consider blocking a complete factorial design with three factors. ***(11 marks)***

4 (i) An opinion poll is to be conducted to estimate the proportion of voters who intend to vote for the Liberal Democrats at the next UK General Election. If a simple random sample is to be used, how large would the sample need to be to ensure that a 99% confidence interval for the true proportion was no wider than 0.05. You may ignore the finite population correction. For your choice of n, would you expect the observed 99% confidence interval to be narrower than 0.05? Explain your answer. ***(8 marks)***

(ii) A survey is conducted to estimate the proportion of current cannabis users in a population. 100 members of the population have been selected using simple random sampling. Each participant first rolls a die, but does not reveal the outcome to the interviewer. If the outcome of rolling the die is a 1, the participant responds "true" or "false" to the following statement

"I have never used cannabis".

If the outcome of rolling the die is 2,3,4,5 or 6, the participant responds "true" or "false" to the following statement

"I have used cannabis at least once".

(a) If r is the proportion of participants who respond "true", derive an unbiased estimator of the proportion of the population who have used cannabis at least once. Calculate the variance of your estimator, as a function of the true proportion of cannabis users.
(7 marks)

(b) Suppose, instead of rolling a die, the participant tosses a coin, with the outcome of heads or tails determining which statement the participant responds to. What would the flaw be with this version of the survey?

(2 marks)

4 (continued)

(iii) A company wishes to estimate the average expenditure on foreign holidays last year by members of a population. The organization has decided to use a stratified random sample with proportional allocation from 3 strata. Information on spending from surveys in the previous year is given in the following table.

Stratum	Size (1,000's)	Est. std. dev.	Est. mean
1	50	800	5000
2	200	300	1000
3	100	100	800

Members of stratum 1 are more difficult to access compared to the other strata, so that there is an extra 20% cost of surveying each individual in stratum 1, compared to strata 2 and 3.

If the aim is to minimise the variance of the estimator of the population mean, within a fixed overall cost of conducting the survey, in what ratio should the three strata be sampled? If, for your chosen ratio, the company calculates that it can afford a total sample size of 1000, calculate the variance of your estimator. ***(13 marks)***

End of Question Paper