Imperial College
London

UNIVERSITY OF LONDON

BSc and MSci EXAMINATIONS (MATHEMATICS)

May-June 2006

This paper is also taken for the relevant examination for the Associateship.

M3S10/M4S10

Design of Experiments and Surveys

Date: Friday, 26th May 2006 Time: 10 am – 12 noon

Credit will be given for all questions attempted but extra credit will be given for complete or nearly complete answers.

Calculators may not be used.

Statistical tables will not be available.

1. (i) Describe two advantages of randomisation in experimental design.

(ii) In the context of a clinical trial to compare the effects of two drugs, A and B, on a disease, describe what is meant by a *double blind trial* and explain the purpose of such a trial.

(iii) In a particular case patients with the disease enter the trial one at a time to see the doctor and each patient is allocated either drug A or drug B. Discuss the problems of randomising the allocation of the two drugs and explain briefly how these problems may be overcome. Illustrate your answer by considering each of the following two scenarios:

(a) Every patient is independently allocated drug A or drug B, each with probability one-half.

(b) If the previous two patients have been allocated drug A then the next patient is allocated drug B. If the previous two patients have been allocated drug B then the next patient is allocated drug A. If the previous two patients were allocated different drugs then the next patient is allocated A or B at random.

2. (i) Define

(a) a balanced incomplete block design (BIBD(t, b, r, k, λ)).

(b) a symmetric BIBD.

(c) a resolvable BIBD.

(ii) By considering a suitable pair of latin squares construct a resolvable balanced incomplete block design D, with $t = 9, b = 12$ and $k = 3$, and explain carefully why your construction works.

(iii) Use design D to construct a symmetric BIBD with $t = 13$ and $k = 4$ and explain carefully why your construction works.

3. (i) In the context of a 2^n factorial experiment explain the difference between *confounding* and *fractional replication*.

(ii) In a particular experiment there are 7 factors A, B, C, D, E, F and G but only 64 experimental units are available and these are arranged in 8 blocks of 8 units each. The 7-factor interaction is aliased with the mean and the following independent effects are confounded with the block effects:

$$ABCG, \quad CDEG, \quad AEFG.$$

Write down the complete set of effects confounded with the block effects.

(iii) Assuming that one of the experimental units receives the treatment combination which has every factor at its low level, write down the other treatment combinations in the same block.

(iv) Suppose instead that the 64 experimental units are arranged in an 8×8 square array. By considering the effects

$$ABFG, \quad ACDG, \quad CEFG$$

show (with justification) how you would allocate the treatment combinations of a $\frac{1}{2}$-replicate of the complete 2^7 factorial experiment, one each, to the cells of this array so that no main effect or two-factor interaction is confounded either with row effects or with column effects. There is no need to write out the whole design but you should include sufficient detail to indicate how this could be done.

4. A response $Y(\underline{x})$ at a point $\underline{x} = (x_1, x_2, \ldots, x_k)^T$ in a closed bounded design region χ of $\mathbb{R}^k$ has the properties

$$\mathsf{E}(Y(\underline{x})) = \underline{f}(\underline{x})^T \underline{\beta} \quad \text{and} \quad \mathsf{var}(Y(\underline{x})) = \sigma^2 > 0,$$

where $\underline{\beta} = (\beta_1, \beta_2, \ldots, \beta_t)^T$ is a $t \times 1$ vector of unknown parameters, $\underline{f}(\underline{x})$ is a $t \times 1$ vector of known continuous functions of $\underline{x}$, and σ^2 is unknown. All observations are independent.

(i) Define (a) the *variance function*, $d(\underline{x}, \xi)$, associated with the design measure ξ over χ and (b) a *G-optimal design measure* ξ^* over χ.

(ii) If the design region χ consists of a finite set of distinct points $\underline{z}_1, \underline{z}_2, \ldots, \underline{z}_m$, show that for any design measure ξ over χ,

$$\sum_{i=1}^{m} d(\underline{z}_i, \xi)\xi(\underline{z}_i) = t.$$

(iii) Deduce that $\max_i d(\underline{z}_i, \xi) \geqslant t$.

(iv) In a particular case $k = 2$, $\chi = \{(2,2), (1,-1), (-1,1)\}$ and

$$\mathsf{E}(Y(\underline{x})) = \beta_1 x_1 + \beta_2 x_2.$$

Consider the design measure ξ_p which attaches weight p to each of $(1,-1)$ and $(-1,1)$ and weight $1-2p$ to $(2,2)$, where $0 < p < \frac{1}{2}$. Find the value of $p\,(= p^*)$ which maximises the determinant of the information matrix of ξ_p.

(v) Show that

$$d(\underline{x}, \xi_{p^*}) = \frac{1}{8}\{5(x_1^2 + x_2^2) - 6x_1x_2\}.$$

(vi) Hence show that ξ_{p^*} is G-optimal over χ.

(vii) Sketch (without proof) the largest region in $\mathbb{R}^2$ over which the design measure ξ_{p^*} is G-optimal.

5. (i) In a sample survey explain what is meant by a *simple random sample* of size n from a population of size N.

(ii) If $\overline{y}$ is the mean of the observations on n sampled units obtained by simple random sampling, write down the formula for var$(\overline{y})$ in terms of n, N and the population variance σ^2.

(iii) Explain what is meant by *stratified random sampling*.

A company manufactures a particular industrial component in each of two factories, one in London and one in Glasgow. The quality of each component can be measured by a characteristic Y and the company wishes to estimate the mean value $\overline{Y}$, of Y, over the two factories, with minimum variance for a fixed total cost C_T. The quality of components is not thought to fluctuate from day to day so they intend taking independent simple random samples of sizes n_1 and n_2 respectively from the outputs of the London and Glasgow factories on a particular day. The selected components are then to be evaluated at the company's laboratory in London. The cost of evaluating any component is $\pounds c$ but the components from the Glasgow factory have to be transported to the laboratory at an extra cost of $\pounds t$ for each component. Transport costs for the components from the London factory can be ignored.

(iv) If on each day the London factory produces twice as many components as the Glasgow factory write down an unbiased estimate $\widetilde{y}$ of $\overline{Y}$ in terms of $\overline{l}$ and $\overline{g}$, the observed sample mean values of Y for the London and Glasgow factories respectively.

(v) Assuming that the population variances for the observations from the two factories are both equal to σ_F^2, show that

$$\text{var}(\widetilde{y}) = \frac{\sigma_F^2}{9}\left(\frac{4}{n_1} + \frac{1}{n_2}\right) + h,$$

where h does not depend on n_1 or n_2.

(vi) Show that providing the daily output of components from the Glasgow factory is sufficiently large, the values of n_1 and n_2 which satisfy the company's requirements are approximately such that

$$\frac{n_1}{n_2} = 2\sqrt{\alpha},$$

where

$$\alpha = 1 + \frac{t}{c}.$$

(vii) Hence show that the optimal values of n_1 and n_2 are approximately

$$n_1 = \frac{2C_T}{c(2+\sqrt{\alpha})} \quad \text{and} \quad n_2 = \frac{C_T}{c(\alpha + 2\sqrt{\alpha})}.$$

(viii) Why must the daily output from the Glasgow factory be sufficiently large?