



The
University
Of
Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS

**Spring Semester
2017–2018**

Sampling Theory and Design of Experiments

2 hours

Candidates may bring to the examination a calculator that 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 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 $x_1 = x_2 = 0$, and the following model is proposed.

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

The investigator proposes the following design (design A) using 4 observations:

Design	Design points
A	$(x_1, x_2) = \{(-1, 1), (-1, 0), (1, 1), (1, -1)\}$

- (i) Are β_1 and β_2 orthogonal to each other in design A? *(3 marks)*
- (ii) If each observation is subject to a measurement error with mean 0 and variance σ^2 , show that $\text{var}(\hat{\beta}_1) < \text{var}(\hat{\beta}_2)$ in design A. *(2 marks)*
- (iii) Sketch the design space with the points in design A clearly indicated. *(2 marks)*
- (iv) With reference to your sketch in (iii) explain why $\text{var}(\hat{\beta}_1) < \text{var}(\hat{\beta}_2)$ in design A. *(3 marks)*
- (v) Justify whether design A is G -optimal. *(5 marks)*
- (vi) A design is called A -optimal if it minimises the sum of the diagonal elements of $(\mathbf{X}^T \mathbf{X})^{-1}$. Consider all **orthogonal** designs for the model $EY = \beta_1 x_1 + \beta_2 x_2$ with 2 design points $(x_1, x_2) \in \{(1, a), (b, c)\}$ such that $-1 \leq x_1, x_2 \leq 1$. Find all A -optimal **orthogonal** designs for this model with 2 design points. *(5 marks)*

- 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
- (2 marks)*
- (b) Let $\alpha_1, \alpha_2, \alpha_3$ and α_4 represent the block parameters for blocks 1, 2, 3 and 4 respectively and let $\beta_1, \beta_2, \beta_3$ and β_4 represent the parameters for teaching methods A, B C and D respectively. Let Y_{ij} represent the response (exam score) in block i using teaching method j . For the design in (a) write down the response vector, the parameter vector and design matrix in full specifying the parameter constraints imposed (assume that the model does not include an intercept term).
- (5 marks)*
- (c) Show how the design in Part (a) could have been obtained from a Latin square.
- (2 marks)*
- (ii) Consider a fractional factorial design with 4 factors (x_1, x_2, x_3, x_4) each of which occurs at two levels, denoted +1 and -1.
- (a) Suppose that four design points are available. Provide two design generators that allow the intercept and the main effects for x_2, x_3 and x_4 to be included in the linear model without confounding. Show the alias structure for these two generators.
- (3 marks)*
- (b) Construct the fractional factorial design using your design in part (ii)(a).
- (3 marks)*
- (iii) Suppose now the interest is only in the two factors x_1 and x_2 .
- (a) Write down the design matrix for a central composite design for x_1 and x_2 with nine observations.
- (3 marks)*
- (b) Justify whether your design would allow you to fit the following linear model

$$E(Y) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_{12}x_1x_2 + \beta_{11}x_1^2 + \beta_{22}x_2^2 + \beta_{111}x_1^3$$

(2 marks)

- 3** Consider taking a simple random sample (SRS) of size 2 from the population X_1, X_2, \dots, X_N with population variance S^2 . Suppose that x_1 and x_2 are obtained and assume further that $E(x_1) = E(x_2) = \bar{X}$ where \bar{X} is the population mean.
- (i) Show that the estimator $\bar{x} = \frac{x_1 + x_2}{2}$ is an unbiased estimator of \bar{X} . *(1 mark)*
 - (ii) Show that $\text{cov}(x_1, x_2) = -\frac{S^2}{N}$. *(5 marks)*
 - (iii) Using the result from (ii) show that $\text{var}(\bar{x}) = \left(1 - \frac{2}{N}\right) \frac{S^2}{2}$. *(4 marks)*
 - (iv) Using the central limit theorem find a 95% CI for \bar{X} assuming that $x_1 = 4.3$ and $x_2 = 3.7$ and that the population is of size 100. *(4 marks)*
 - (v) Justify whether your confidence interval in part (iii) is reliable or not. *(1 mark)*
 - (vi) Suppose that in an SRS of size 2 only x_1^2 and x_2^2 are available instead of x_1 and x_2 . Instead of using an estimator of \bar{X} an investigator proposes an estimator of \bar{X}^2 . They propose the following estimator $x_{se} = \frac{x_1^2 + x_2^2}{k}$ for some $k \in \mathbb{R}^+, k \neq 2$ and $2 \leq N$. Discuss when x_{se} is unbiased for \bar{X}^2 and whether this would be a sensible estimator in practice. *(5 marks)*

End of Question Paper

MAS370 FORMULAE & CRITICAL VALUES)

1 Design Formulae

Linear Model formulae

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} \quad \text{and} \quad \hat{\beta} \sim N\{\beta, \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}\}$$

Prediction Variance

$$\text{var } \hat{y}(x_0) = \sigma^2 \mathbf{f}(x_0)^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{f}(x_0)$$

Standardized Prediction Variance

$$d(\mathbf{x}) = n \mathbf{f}(\mathbf{x})^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{f}(\mathbf{x}) = \mathbf{f}(\mathbf{x})^T \mathbf{M}^{-1} \mathbf{f}(\mathbf{x})$$

Confidence Regions, σ^2 unknown

$$p^{-1} \hat{\sigma}^{-2} (\hat{\beta} - \beta)^T \mathbf{X}^T \mathbf{X} (\hat{\beta} - \beta) \text{ has an } F_{p, n-p} \text{ distribution, provided } n > p$$

Balanced Incomplete Block Design Notation

- t = number of treatments
- k = number of units in a block
- b = number of blocks
- r = number of applications of each treatment
- λ = number of times each pair of treatments appears together in a block

Balanced Incomplete Block Design Relationships

$$\begin{aligned} t &> k \\ bk &= rt \\ r(k-1) &= \lambda(t-1) \end{aligned}$$

Balanced Incomplete Block Design - Unreduced Design

$$b = \binom{t}{k} \quad r = \binom{t-1}{k-1} \quad \lambda = \binom{t-2}{k-2}$$

Efficiency of Balanced Incomplete Block Design compared to a Randomized Block design

$$\frac{1 - t^{-1}}{1 - k^{-1}}$$

Adding an extra point x

$$|\mathbf{G}^*| = |\mathbf{G}| (1 + \mathbf{f}(x)^T \mathbf{G}^{-1} \mathbf{f}(x))$$

Deleting an existing point x

$$|\mathbf{G}^*| = |\mathbf{G}| (1 - \mathbf{f}(x)^T \mathbf{G}^{-1} \mathbf{f}(x))$$

Adding a new point y and deleting an existing point x

$$|\mathbf{G}_2| = |\mathbf{G}| \left\{ (1 - \mathbf{f}(x)^T \mathbf{G}^{-1} \mathbf{f}(x)) (1 + \mathbf{f}(y)^T \mathbf{G}^{-1} \mathbf{f}(y)) + (\mathbf{f}(x)^T \mathbf{G}^{-1} \mathbf{f}(y))^2 \right\}$$

2 Sample Surveys and Computer Experiments Formulae

Population variance

$$S^2 = \frac{1}{N-1} \sum_1^N (X_i - \bar{X})^2 = \frac{1}{N-1} \left(\sum_{i=1}^N X_i^2 - N\bar{X}^2 \right)$$

and for a binary characteristic ($X_i = 1$ or 0 for each i),

$$S^2 = \frac{NP(1-P)}{N-1}$$

Variance of sample mean for simple random sampling

$$\text{var } \bar{x} = \left(1 - \frac{n}{N}\right) \frac{S^2}{n}$$

Sample size to achieve given confidence interval width for simple random sampling

$$n \geq \frac{N}{1 + N(d/(2Sz_{\alpha/2}))^2}$$

Stratified estimate of population mean and its variance

$$\bar{x}_{st} = \frac{1}{N} \sum_1^l N_i \bar{x}_i \quad \text{and} \quad \text{var } \bar{x}_{st} = \sum_1^l \left(\frac{N_i}{N}\right)^2 \frac{1-f_i}{n_i} S_i^2.$$

Optimal allocation

$$n_i \propto \frac{N_i S_i}{\sqrt{c_i}}$$

Neyman allocation

$$n_i = \frac{n N_i S_i}{\sum_1^l N_i S_i}$$

Cluster estimate of population mean and its variance

$$\bar{x}_{cl} = \frac{1}{lK} \sum_1^l \sum_1^K x_{ij} \quad \text{and} \quad \text{var } (\bar{x}_{cl}) = \frac{1-f}{l} \frac{1}{L-1} \sum_1^L (\bar{X}_i - \bar{X})^2$$

Regression estimator of population mean and its variance

$$\bar{x}_{lr} = \bar{x} - \hat{\beta}(\bar{y} - \bar{Y}) \quad \text{and} \quad \text{var } \bar{x}_{lr} \simeq \frac{1-f}{n} S_X^2 (1 - \rho^2)$$

Approximate variance of the Peterson estimator, Chapman estimator and approximate variance

n : size of 1st sample, m : size of 2nd sample.

$$\begin{aligned} \widehat{Var}(\hat{N}_p) &= \frac{mn^2(m-r)}{r^3}, \\ \hat{N}_c &= \frac{(n+1)(m+1)}{r+1} - 1, \\ \widehat{Var}(\hat{N}_c) &= \frac{(n+1)(m+1)(n-r)(m-r)}{(r+1)^2(r+2)}. \end{aligned}$$

Variance identity

$$\text{Var}(Y) = \text{Var}_X\{E(Y|X)\} + E_X\{\text{Var}(Y|X)\}.$$

3 Tables of Percentage Points (also known as Quantiles or Critical Values) for Three Standard Distributions

The tables contain values of quantiles q such that $P[X \leq q] = p$ for various probabilities p when X has the specified distribution (which may depend on particular degrees of freedom ν). In these tables, p has been expressed as a percentage rather than a decimal. The relevant R commands for generating the q are also shown. For the $N(0, 1)$ distribution, the tabulated function is also known as the Φ^{-1} function.

STANDARD NORMAL DISTRIBUTION PERCENTAGE POINTS

`qnorm(p)` where p is percentage, e.g. for 95%, $p = 0.95$

	60.0%	66.7%	75.0%	80.0%	87.5%	90.0%	95.0%	97.5%	99.0%	99.5%	99.9%
<code>qnorm</code>	0.253	0.431	0.674	0.842	1.150	1.282	1.645	1.960	2.326	2.576	3.090

CHI-SQUARED PERCENTAGE POINTS

`qchisq(p, nu)` where p is percentage, e.g. for 95%, $p = 0.95$

ν	60.0%	66.7%	75.0%	80.0%	87.5%	90.0%	95.0%	97.5%	99.0%	99.5%	99.9%
1	0.708	0.936	1.323	1.642	2.354	2.706	3.841	5.024	6.635	7.879	10.828
2	1.833	2.197	2.773	3.219	4.159	4.605	5.991	7.378	9.210	10.597	13.816
3	2.946	3.405	4.108	4.642	5.739	6.251	7.815	9.348	11.345	12.838	16.266
4	4.045	4.579	5.385	5.989	7.214	7.779	9.488	11.143	13.277	14.860	18.467
5	5.132	5.730	6.626	7.289	8.625	9.236	11.070	12.833	15.086	16.750	20.515
6	6.211	6.867	7.841	8.558	9.992	10.645	12.592	14.449	16.812	18.548	22.458
7	7.283	7.992	9.037	9.803	11.326	12.017	14.067	16.013	18.475	20.278	24.322
8	8.351	9.107	10.219	11.030	12.636	13.362	15.507	17.535	20.090	21.955	26.125
9	9.414	10.215	11.389	12.242	13.926	14.684	16.919	19.023	21.666	23.589	27.877
10	10.473	11.317	12.549	13.442	15.198	15.987	18.307	20.483	23.209	25.188	29.588

STUDENT'S t PERCENTAGE POINTS
 $qt(p, \nu)$ where p is percentage, e.g. for 95%, $p = 0.95$

ν	60.0%	66.7%	75.0%	80.0%	87.5%	90.0%	95.0%	97.5%	99.0%	99.5%	99.9%
1	0.325	0.577	1.000	1.376	2.414	3.078	6.314	12.706	31.821	63.657	318.31
2	0.289	0.500	0.816	1.061	1.604	1.886	2.920	4.303	6.965	9.925	22.327
3	0.277	0.476	0.765	0.978	1.423	1.638	2.353	3.182	4.541	5.841	10.215
4	0.271	0.464	0.741	0.941	1.344	1.533	2.132	2.776	3.747	4.604	7.173
5	0.267	0.457	0.727	0.920	1.301	1.476	2.015	2.571	3.365	4.032	5.893
6	0.265	0.453	0.718	0.906	1.273	1.440	1.943	2.447	3.143	3.707	5.208
7	0.263	0.449	0.711	0.896	1.254	1.415	1.895	2.365	2.998	3.499	4.785
8	0.262	0.447	0.706	0.889	1.240	1.397	1.860	2.306	2.896	3.355	4.501
9	0.261	0.445	0.703	0.883	1.230	1.383	1.833	2.262	2.821	3.250	4.297
10	0.260	0.444	0.700	0.879	1.221	1.372	1.812	2.228	2.764	3.169	4.144
11	0.260	0.443	0.697	0.876	1.214	1.363	1.796	2.201	2.718	3.106	4.025
12	0.259	0.442	0.695	0.873	1.209	1.356	1.782	2.179	2.681	3.055	3.930
13	0.259	0.441	0.694	0.870	1.204	1.350	1.771	2.160	2.650	3.012	3.852
14	0.258	0.440	0.692	0.868	1.200	1.345	1.761	2.145	2.624	2.977	3.787
15	0.258	0.439	0.691	0.866	1.197	1.341	1.753	2.131	2.602	2.947	3.733
16	0.258	0.439	0.690	0.865	1.194	1.337	1.746	2.120	2.583	2.921	3.686
17	0.257	0.438	0.689	0.863	1.191	1.333	1.740	2.110	2.567	2.898	3.646
18	0.257	0.438	0.688	0.862	1.189	1.330	1.734	2.101	2.552	2.878	3.610
19	0.257	0.438	0.688	0.861	1.187	1.328	1.729	2.093	2.539	2.861	3.579
20	0.257	0.437	0.687	0.860	1.185	1.325	1.725	2.086	2.528	2.845	3.552
21	0.257	0.437	0.686	0.859	1.183	1.323	1.721	2.080	2.518	2.831	3.527
22	0.256	0.437	0.686	0.858	1.182	1.321	1.717	2.074	2.508	2.819	3.505
23	0.256	0.436	0.685	0.858	1.180	1.319	1.714	2.069	2.500	2.807	3.485
24	0.256	0.436	0.685	0.857	1.179	1.318	1.711	2.064	2.492	2.797	3.467
25	0.256	0.436	0.684	0.856	1.178	1.316	1.708	2.060	2.485	2.787	3.450
26	0.256	0.436	0.684	0.856	1.177	1.315	1.706	2.056	2.479	2.779	3.435
27	0.256	0.435	0.684	0.855	1.176	1.314	1.703	2.052	2.473	2.771	3.421
28	0.256	0.435	0.683	0.855	1.175	1.313	1.701	2.048	2.467	2.763	3.408
29	0.256	0.435	0.683	0.854	1.174	1.311	1.699	2.045	2.462	2.756	3.396
30	0.256	0.435	0.683	0.854	1.173	1.310	1.697	2.042	2.457	2.750	3.385
35	0.255	0.434	0.682	0.852	1.170	1.306	1.690	2.030	2.438	2.724	3.340
40	0.255	0.434	0.681	0.851	1.167	1.303	1.684	2.021	2.423	2.704	3.307
45	0.255	0.434	0.680	0.850	1.165	1.301	1.679	2.014	2.412	2.690	3.281
50	0.255	0.433	0.679	0.849	1.164	1.299	1.676	2.009	2.403	2.678	3.261
55	0.255	0.433	0.679	0.848	1.163	1.297	1.673	2.004	2.396	2.668	3.245
60	0.254	0.433	0.679	0.848	1.162	1.296	1.671	2.000	2.390	2.660	3.232
∞	0.253	0.431	0.674	0.842	1.150	1.282	1.645	1.960	2.326	2.576	3.090