SCHOOL OF MATHEMATICS AND STATISTICS

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:

[ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ]
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. It is assumed that $E(Y) = 0$ when $x = 0$, and the following model (model 1) is proposed.

$$E(Y) = \beta_1 x + \beta_{11} x^2.$$  

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

<table>
<thead>
<tr>
<th>Design</th>
<th>Design points $(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>${-1, 0, 0, 1}$</td>
</tr>
</tbody>
</table>

(i) Explain why $\beta_1$ and $\beta_{11}$ in Model 1 are orthogonal to each other in design A.  
 *(3 marks)*

(ii) The investigator thinks that collecting further observations at $x = 0$ would reduce the prediction variance for all $x$ satisfying $-1 \leq x \leq 1$ for design A in model 1. Justify whether they are correct or not.  
 *(5 marks)*

(iii) Suppose the design aim is to minimise the variance of $\hat{\beta}_{11}$ in model 1 and that only four design points can be used. Specify the four design points that achieve this aim and that also ensure that $\beta_1$ is orthogonal to $\beta_{11}$. Justify your answer.  
 *(2 marks)*

(iv) Suppose instead the investigator proposes the following model (model 2) 

$$E(Y) = \beta_1 x + \beta_{11} x^2 + \beta_{111} x^3.$$  

Justify whether design A is a suitable design for model 2.  
 *(7 marks)*

(v) The investigator now believes an intercept is needed and proposes another new model (model 3) 

$$E(Y) = \beta + \beta_1 x + \beta_{11} x^2.$$  

Justify whether design A is $G$-optimal for model 3. You may find the following result useful

$$\begin{pmatrix} 2 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & 0 \\ -1 & 0 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$  
 *(7 marks)*
Throughout this question assume that all four variables \((x_1, x_2, x_3\) and \(x_4\)) are factors occurring at two levels, denoted +1 and −1.

(i) Suppose that four design points are available in a fractional factorial design. Specify two design generators that allow the intercept and the main effects for \(x_1, x_2\) and \(x_3\) to be included in a linear model without confounding. Show the alias structure for these two generators (you can leave the alias structure in terms of \(x_1, x_2, x_3\) and \(x_4\) rather than in terms of \(\beta_{1}, \beta_{2}, \ldots, \beta_{1234}\)).

\((3\ \text{marks})\)

(ii) Construct the fractional factorial design using your design in part (i).

\((3\ \text{marks})\)

(iii) Consider a screening experiment with four variables \(x_1, x_2, x_3, x_4\). If the interest is only in the main effects of \(x_1, x_2, x_3\) and \(x_4\), construct a design suitable for a screening experiment using eight design points.

\((5\ \text{marks})\)

(iv) Write down a Latin square that could be used as a design with 16 design points for an experiment with three factor variables each taking four levels.

\((2\ \text{marks})\)

(v) Suppose the investigator also wants to include a blocking factor with four levels into the design in part (iv). Describe how this could be done with 16 design points.

\((3\ \text{marks})\)

(vi) Write down the linear model, that includes only the main effects of the four factors, that could be used with the design in part (v). Define all notation and specify all constraints.

\((4\ \text{marks})\)
A small survey has been conducted to estimate the proportion of the population in favour of lowering the retirement age. The sample was drawn using simple random sampling. The sex of each participant in the survey was also recorded, and the results are given below.

<table>
<thead>
<tr>
<th></th>
<th>males</th>
<th>females</th>
</tr>
</thead>
<tbody>
<tr>
<td>in favour</td>
<td>23</td>
<td>15</td>
</tr>
<tr>
<td>against</td>
<td>18</td>
<td>44</td>
</tr>
</tbody>
</table>

(i) Estimate the population proportion in favour of lowering the retirement age using the simple random sampling estimator. 
\[ \text{(2 marks)} \]

(ii) Assume that half the population are female. Estimate the population proportion in favour of lowering the retirement age using an alternative estimator that makes use of the assumed proportion of females in the population. 
\[ \text{(2 marks)} \]

(iii) 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. 
\[ \text{(5 marks)} \]

(iv) For your choice of \( n \) in part (iii), would you expect the observed 99% confidence interval to be narrower than 0.05? Explain your answer. 
\[ \text{(3 marks)} \]

(v) Discuss briefly when cluster sampling might give a less accurate estimator of the population mean than using simple random sampling (drawing a diagram here might be helpful). 
\[ \text{(3 marks)} \]

(vi) In cluster sampling that we studied, the cluster sizes are assumed to be equal to each other. Suppose instead that this is no longer true. Specifically, suppose that there are three clusters of sizes \( k_1, k_2 \) and \( k_3 \) where it is not true that \( k_1 = k_2 = k_3 \). Consider an estimator that is the sample mean of two randomly selected clusters (assume that you still measure all units in the selected clusters). Justify whether this is an unbiased estimator of the population mean. 
\[ \text{(5 marks)} \]

End of Question Paper
1 Design Formulae

Linear Model formulae

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

Prediction Variance

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

Standardized Prediction Variance

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

Confidence Regions, \( \sigma^2 \) unknown

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

Balanced Incomplete Block Design Notation

\[
t = \text{number of treatments} \\
k = \text{number of units in a block} \\
b = \text{number of blocks} \\
r = \text{number of applications of each treatment} \\
\lambda = \text{number of times each pair of treatments appears together in a block}
\]

Balanced Incomplete Block Design Relationships

\[
t > k \\
bk = rt \\
r(k-1) = \lambda(t-1)
\]

Balanced Incomplete Block Design - Unreduced Design

\[
b = \left( \begin{array}{c} t \\ k \end{array} \right) \\
r = \left( \frac{t-1}{k-1} \right) \\
\lambda = \left( \frac{t-2}{k-2} \right)
\]

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

\[
\frac{1 - t^{-1}}{1 - k^{-1}}
\]

Adding an extra point \( x \)

\[
|G^*| = |G| \left(1 + f(x)^T G^{-1} f(x)\right)
\]

Deleting an existing point \( x \)

\[
|G^*| = |G| \left(1 - f(x)^T G^{-1} f(x)\right)
\]

Adding a new point \( y \) and deleting an existing point \( x \)

\[
|G_2| = |G| \left\{ (1 - f(x)^T G^{-1} f(x)) \left(1 + f(y)^T G^{-1} f(y)\right) + (f(x)^T G^{-1} f(y))^2 \right\}
\]
2 Sample Surveys and Computer Experiments Formulae

Population variance

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

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

\[ n \geq \frac{N}{1 + N(d/(2S\sigma/2))^2} \]

Stratified estimate of population mean and its variance

\[ \bar{x}_{st} = \frac{1}{N} \sum_{i=1}^{I} N_i \bar{x}_i \quad \text{and} \quad \text{var } \bar{x}_{st} = \sum_{i=1}^{I} \left( \frac{N_i}{N} \right)^2 \frac{1-f_i}{n_i} S_i^2 \]

Optimal allocation

\[ n_i \propto \frac{N_i S_i}{\sqrt{c_i}} \]

Neyman allocation

\[ n_i = \frac{n N_i S_i}{\sum_{i=1}^{I} N_i S_i} \]

Cluster estimate of population mean and its variance

\[ \bar{x}_{cl} = \frac{1}{K} \sum_{i=1}^{I} \sum_{K} x_{ij} \quad \text{and} \quad \text{var } \bar{x}_{cl} = \frac{1-f}{L-1} \sum_{i=1}^{I} (\bar{X}_i - \bar{X})^2 \]

Regression estimator of population mean and its variance

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

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

\[ \hat{N}_p = \frac{mn^2(m-r)}{r^2}, \]

\[ \hat{N}_c = \frac{(n+1)(m+1)}{r+1} - 1, \]

\[ \hat{V}ar(\hat{N}_c) = \frac{(n+1)(m+1)(n-r)(m-r)}{(r+1)^2(r+2)}. \]

Variance identity

\[ Var(Y) = Var \{E(Y|X)\} + E \{Var(Y|X)\}. \]
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 $\nu$). 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 $\Phi^{-1}$ function.

**STANDARD NORMAL DISTRIBUTION PERCENTAGE POINTS**

$q_{\text{norm}}(p)$ where $p$ is percentage, e.g. for 95%, $p = 0.95$

<table>
<thead>
<tr>
<th>$\nu$</th>
<th>60.0%</th>
<th>66.7%</th>
<th>75.0%</th>
<th>80.0%</th>
<th>87.5%</th>
<th>90.0%</th>
<th>95.0%</th>
<th>97.5%</th>
<th>99.0%</th>
<th>99.5%</th>
<th>99.9%</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.708</td>
<td>0.936</td>
<td>1.323</td>
<td>1.642</td>
<td>2.354</td>
<td>2.706</td>
<td>3.841</td>
<td>5.024</td>
<td>6.635</td>
<td>7.879</td>
<td>10.828</td>
</tr>
</tbody>
</table>

**CHI-SQUARED PERCENTAGE POINTS**

$q_{\text{chisq}}(p, \nu)$ where $p$ is percentage, e.g. for 95%, $p = 0.95$

<table>
<thead>
<tr>
<th>$\nu$</th>
<th>60.0%</th>
<th>66.7%</th>
<th>75.0%</th>
<th>80.0%</th>
<th>87.5%</th>
<th>90.0%</th>
<th>95.0%</th>
<th>97.5%</th>
<th>99.0%</th>
<th>99.5%</th>
<th>99.9%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.708</td>
<td>0.936</td>
<td>1.323</td>
<td>1.642</td>
<td>2.354</td>
<td>2.706</td>
<td>3.841</td>
<td>5.024</td>
<td>6.635</td>
<td>7.879</td>
<td>10.828</td>
</tr>
</tbody>
</table>
### STUDENT'S t PERCENTAGE POINTS

qt(p, ν) where p is percentage, e.g. for 95%, p = 0.95

<table>
<thead>
<tr>
<th>ν</th>
<th>60.0%</th>
<th>66.7%</th>
<th>75.0%</th>
<th>80.0%</th>
<th>87.5%</th>
<th>90.0%</th>
<th>95.0%</th>
<th>97.5%</th>
<th>99.0%</th>
<th>99.5%</th>
<th>99.9%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.325</td>
<td>0.577</td>
<td>1.000</td>
<td>1.376</td>
<td>2.414</td>
<td>3.078</td>
<td>6.314</td>
<td>12.706</td>
<td>31.821</td>
<td>63.657</td>
<td>318.31</td>
</tr>
<tr>
<td>2</td>
<td>0.289</td>
<td>0.450</td>
<td>0.816</td>
<td>1.061</td>
<td>1.604</td>
<td>1.866</td>
<td>3.841</td>
<td>5.598</td>
<td>7.173</td>
<td>8.696</td>
<td>9.925</td>
</tr>
<tr>
<td>3</td>
<td>0.277</td>
<td>0.403</td>
<td>0.765</td>
<td>0.978</td>
<td>1.341</td>
<td>1.638</td>
<td>2.920</td>
<td>4.303</td>
<td>5.240</td>
<td>6.430</td>
<td>7.281</td>
</tr>
<tr>
<td>4</td>
<td>0.271</td>
<td>0.392</td>
<td>0.741</td>
<td>0.941</td>
<td>1.346</td>
<td>1.675</td>
<td>2.841</td>
<td>4.032</td>
<td>4.998</td>
<td>5.893</td>
<td>6.945</td>
</tr>
<tr>
<td>5</td>
<td>0.267</td>
<td>0.385</td>
<td>0.727</td>
<td>0.920</td>
<td>1.353</td>
<td>1.631</td>
<td>2.763</td>
<td>3.774</td>
<td>4.604</td>
<td>5.407</td>
<td>5.999</td>
</tr>
<tr>
<td>6</td>
<td>0.265</td>
<td>0.380</td>
<td>0.718</td>
<td>0.906</td>
<td>1.363</td>
<td>1.571</td>
<td>2.681</td>
<td>3.552</td>
<td>4.297</td>
<td>5.041</td>
<td>5.598</td>
</tr>
<tr>
<td>7</td>
<td>0.263</td>
<td>0.376</td>
<td>0.711</td>
<td>0.896</td>
<td>1.373</td>
<td>1.508</td>
<td>2.603</td>
<td>3.355</td>
<td>3.930</td>
<td>4.587</td>
<td>5.144</td>
</tr>
<tr>
<td>8</td>
<td>0.262</td>
<td>0.373</td>
<td>0.706</td>
<td>0.889</td>
<td>1.383</td>
<td>1.486</td>
<td>2.535</td>
<td>3.143</td>
<td>3.611</td>
<td>4.221</td>
<td>4.770</td>
</tr>
<tr>
<td>9</td>
<td>0.261</td>
<td>0.370</td>
<td>0.702</td>
<td>0.884</td>
<td>1.394</td>
<td>1.469</td>
<td>2.473</td>
<td>2.950</td>
<td>3.341</td>
<td>3.907</td>
<td>4.412</td>
</tr>
<tr>
<td>10</td>
<td>0.260</td>
<td>0.367</td>
<td>0.699</td>
<td>0.879</td>
<td>1.405</td>
<td>1.452</td>
<td>2.417</td>
<td>2.781</td>
<td>3.122</td>
<td>3.601</td>
<td>4.115</td>
</tr>
<tr>
<td>11</td>
<td>0.260</td>
<td>0.365</td>
<td>0.696</td>
<td>0.875</td>
<td>1.416</td>
<td>1.438</td>
<td>2.364</td>
<td>2.632</td>
<td>2.937</td>
<td>3.356</td>
<td>3.852</td>
</tr>
<tr>
<td>12</td>
<td>0.259</td>
<td>0.363</td>
<td>0.693</td>
<td>0.872</td>
<td>1.427</td>
<td>1.425</td>
<td>2.317</td>
<td>2.508</td>
<td>2.784</td>
<td>3.159</td>
<td>3.625</td>
</tr>
<tr>
<td>13</td>
<td>0.259</td>
<td>0.361</td>
<td>0.691</td>
<td>0.869</td>
<td>1.439</td>
<td>1.413</td>
<td>2.274</td>
<td>2.390</td>
<td>2.660</td>
<td>2.998</td>
<td>3.427</td>
</tr>
<tr>
<td>14</td>
<td>0.258</td>
<td>0.359</td>
<td>0.688</td>
<td>0.866</td>
<td>1.451</td>
<td>1.403</td>
<td>2.235</td>
<td>2.279</td>
<td>2.552</td>
<td>2.821</td>
<td>3.251</td>
</tr>
<tr>
<td>15</td>
<td>0.258</td>
<td>0.358</td>
<td>0.686</td>
<td>0.864</td>
<td>1.464</td>
<td>1.394</td>
<td>2.201</td>
<td>2.179</td>
<td>2.457</td>
<td>2.681</td>
<td>3.106</td>
</tr>
<tr>
<td>20</td>
<td>0.257</td>
<td>0.354</td>
<td>0.678</td>
<td>0.855</td>
<td>1.400</td>
<td>1.333</td>
<td>1.740</td>
<td>2.101</td>
<td>2.552</td>
<td>2.878</td>
<td>3.610</td>
</tr>
<tr>
<td>25</td>
<td>0.256</td>
<td>0.350</td>
<td>0.671</td>
<td>0.849</td>
<td>1.370</td>
<td>1.294</td>
<td>1.692</td>
<td>2.056</td>
<td>2.485</td>
<td>2.779</td>
<td>3.435</td>
</tr>
<tr>
<td>30</td>
<td>0.255</td>
<td>0.346</td>
<td>0.664</td>
<td>0.844</td>
<td>1.344</td>
<td>1.264</td>
<td>1.653</td>
<td>2.021</td>
<td>2.423</td>
<td>2.690</td>
<td>3.300</td>
</tr>
<tr>
<td>40</td>
<td>0.254</td>
<td>0.338</td>
<td>0.652</td>
<td>0.830</td>
<td>1.299</td>
<td>1.201</td>
<td>1.554</td>
<td>1.930</td>
<td>2.326</td>
<td>2.576</td>
<td>3.090</td>
</tr>
<tr>
<td>50</td>
<td>0.253</td>
<td>0.331</td>
<td>0.645</td>
<td>0.822</td>
<td>1.267</td>
<td>1.151</td>
<td>1.494</td>
<td>1.856</td>
<td>2.232</td>
<td>2.518</td>
<td>2.947</td>
</tr>
<tr>
<td>60</td>
<td>0.253</td>
<td>0.327</td>
<td>0.639</td>
<td>0.816</td>
<td>1.242</td>
<td>1.114</td>
<td>1.454</td>
<td>1.791</td>
<td>2.150</td>
<td>2.483</td>
<td>2.831</td>
</tr>
<tr>
<td>∞</td>
<td>0.252</td>
<td>0.323</td>
<td>0.634</td>
<td>0.810</td>
<td>1.220</td>
<td>1.081</td>
<td>1.421</td>
<td>1.738</td>
<td>2.080</td>
<td>2.452</td>
<td>2.734</td>
</tr>
</tbody>
</table>