UNIVERSITY OF CAMBRIDGE

M. P*HIL. IN* STATISTICAL SCIENCE

Monday 3 June 2002 9 to 12

EXPERIMENTAL DESIGN AND MULTIVARIATE ANALYSIS

*Attempt **FOUR** questions*

*There are **six** questions in total*

The questions carry equal weight

You may not start to read the questions printed on the subsequent pages until instructed to do so by the Invigilator.

1 Applied Multivariate Analysis

(i) Given independent random vectors $Y_1, \ldots, Y_n$ from the p-variate distribution $N(\mu, V)$, with μ and V unknown, prove that the maximum likelihood estimates of μ, V are

$$\bar{Y} = \frac{1}{n}\sum_1^n Y_i, \quad S = \frac{1}{n}\sum_1^n (Y_i - \bar{Y})(Y_i - \bar{Y})^T.$$

(ii) With $Y_1, \ldots, Y_n$ as above, show that there exists a $p \times p$ matrix L such that

$$L(Y_i - \mu) \sim N(0, I) \quad \text{for each } i,\ 1 \leqslant i \leqslant n\ .$$

Hence show that the distribution of

$$(\bar{Y} - \mu)^T\ S^{-1}(\bar{Y} - \mu)$$

is free of μ, V.

2 Applied Multivariate Analysis

The R data-set US Judge Ratings contains lawyers' ratings of 43 different judges, according to 11 numeric variables, these being

DMNR = Demeanour
DILG = Diligence
CFMG = Case flow managing
DECI = Prompt decisions
PREP = Preparation for trial
FAMI = Familiarity with law
ORAL = Sound oral rulings
WRIT = Sound written ruling
PHYS = Physical ability
RTEN = Worthy of retention.

Interpret the (slightly edited) SPlus5-output given below, explaining carefully the terms that you use, and describe how a 2-dimensional representation of the 43 judges may be constructed from these data.

[*You may assume that the data-set is in the* 43×11 *matrix 'judges'.*]

```
>judges.pc _ princomp(judges, cor=T) ; summary(judges.pc)
Importance of components:
                          Comp. 1    Comp. 2    Comp. 3     Comp. 4     Comp. 5
    Standard deviation 3.1833029 0.65163561 0.50525195 0.302163952 0.193132512
Proportion of Variance 0.9212198 0.03860263 0.02320723 0.008300278 0.003390924
 Cumulative Proportion 0.9212198 0.95982240 0.98302963 0.991329907 0.994720832
.......................
.......................
```

```
>loadings(judges.pc)

     Comp. 1 Comp. 2
INTG  0.289   0.574
DMNR  0.287   0.576
DILG  0.304  -0.139
CFMG  0.303  -0.310
DECI  0.302  -0.336
PREP  0.309  -0.125
FAMI  0.307  -0.123
ORAL  0.313
WRIT  0.311
PHYS  0.281  -0.235
RTEN  0.310   0.153
.............................
```

3 Applied Multivariate Analysis

Fisher's classic "iris" data consists of a table 150×5, of which the first 3 rows are given in the SPlus5 output below. There are 3 distinct species, denoted here by "s", "c" and "v", and we wish to construct a classification tree to sort the 150 iris specimens into species according to the values of the Sepal Length, Sepal Width, Petal Length and Petal Width. Explain carefully the construction of the SPlus object "ir.tr" in the output below, and sketch the resulting classification tree.

```
 > ird[1:3,]
  Sepal.L. Sepal.W. Petal.L. Petal.W. Species
1      5.1      3.5      1.4      0.2       s
2      4.9      3.0      1.4      0.2       s
3      4.7      3.2      1.3      0.2       s
> ir.tr _ tree(Species ~ . , ird) ; summary(ir.tr)

Classification tree:
tree(formula = Species ~ ., data = ird)
Variables actually used in tree construction:
[1] "Petal.L." "Petal.W." "Sepal.L."
Number of terminal nodes:  6
Residual mean deviance:  0.1253 = 18.05 / 144
Misclassification error rate: 0.02667 = 4 / 150
>ir.tr
node), split, n, deviance, yval, (yprob)
      * denotes terminal node

 1) root 150 329.600 c ( 0.33330 0.3333 0.33330 )
   2) Petal.L.<2.45 50   0.000 s ( 0.00000 1.0000 0.00000 ) *
   3) Petal.L.>2.45 100 138.600 c ( 0.50000 0.0000 0.50000 )
     6) Petal.W.<1.75 54  33.320 c ( 0.90740 0.0000 0.09259 )
      12) Petal.L.<4.95 48   9.721 c ( 0.97920 0.0000 0.02083 )
        24) Sepal.L.<5.15 5   5.004 c ( 0.80000 0.0000 0.20000 ) *
        25) Sepal.L.>5.15 43   0.000 c ( 1.00000 0.0000 0.00000 ) *
      13) Petal.L.>4.95 6   7.638 v ( 0.33330 0.0000 0.66670 ) *
     7) Petal.W.>1.75 46   9.635 v ( 0.02174 0.0000 0.97830 )
      14) Petal.L.<4.95 6   5.407 v ( 0.16670 0.0000 0.83330 ) *
      15) Petal.L.>4.95 40   0.000 v ( 0.00000 0.0000 1.00000 ) *
```

4 Design of Experiments

Describe what is meant by a *randomised complete block design* and a *balanced incomplete block design*. Explain what is meant by a *Latin square* and a *Graeco-Latin square*.

An experiment was carried out to investigate operators of a certain type of hydrostatic testing machine used to test the water repellency of fabrics. Because water repellency is thought to vary along the length and width of the fabric, the available square of fabric was divided into four equal parts along the width and four equal parts along the length, making sixteen smaller squares in all. These were used in the Latin square design below, where $1, 2, 3, 4$ denote the four operators

$$\begin{array}{cccc} 2 & 4 & 3 & 1 \\ 1 & 3 & 4 & 2 \\ 4 & 2 & 1 & 3 \\ 3 & 1 & 2 & 4 \end{array}$$

The following analysis of variance table was obtained

Source	*df*	Sum of Squares
row	$*$	5.00
column	$*$	8.50
operator	$*$	18.25
residual	$*$	17.25
total (corrected)	$*$	49.00

Complete the missing degrees of freedom. Write down the model that has been fitted, and test whether there is an operator effect. Would your answer be changed if row, column and operator had been fitted in a different order?

In fact there were four machines, A, B, C and D, and the design used was

$$\begin{array}{cccc} 2B & 4A & 3D & 1C \\ 1A & 3B & 4C & 2D \\ 4D & 2C & 1B & 3A \\ 3C & 1D & 2A & 4B \end{array}$$

What sort of design is this? Fitting the appropriate model gives the sum of squares for machines as 16.50. Now test for an operator effect.

$$\begin{array}{llll} [& F_{4,3}\ (0.05) = 9.12 & F_{3,6}(0.05) = 4.76 & F_{6,3}(0.05) = \ \ 6.16 \\ & F_{3,15}(0.05) = 3.29 & F_{3,3}(0.05) = 9.28 & F_{3,3}(0.01) = 29.46 \quad] \end{array}$$

5 Design of Experiments

Explain how one contrast can divide the treatments in a 2^m design into two blocks each of size 2^{m-1}. What does it mean to say that this contrast is *confounded* with blocks?

An experiment is carried out to investigate the yield from a chemical process using two different catalysts, two different temperatures and two different pressures. One day's blend of the input chemical is just enough for four runs, so that the experiment takes place over two days as follows:-

temperature	pressure	catalyst	day
low	low	1	1
high	low	1	2
low	high	1	2
high	high	1	1
low	low	2	2
high	low	2	1
low	high	2	1
high	high	2	2

Treating days as blocks, determine which contrast has been confounded with days. If all two-factor and higher order interactions are assumed negligible, what is the partition of the degrees of freedom for the resulting analysis of variance?

A second chemical process with the same catalysts, and the same levels of temperature and pressure, is to be investigated using a different procedure such that one day's blend is only enough for two runs. It is possible to arrange the runs so that all main effects can be estimated? It is possible to arrange the runs so that all main effects and the temperature-catalyst interaction can be estimated? In both cases, if it is possible, then given an appropriate design, and if it is not possible, then explain why.

6 Design of Experiments

Consider the model $y = (f_1(\mathbf{x}), \ldots, f_p(\mathbf{x}))\boldsymbol{\beta} + \epsilon$ where $\mathbf{x} = (x_1, x_2) \in [-1, 1]^2$.

Explain what is meant by a *D-optimal* design and by a *G-optimal* design. State the General Equivalence Theorem, defining any terms that you use.

Suppose $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$ with $\mathbb{E}(\epsilon) = 0$ and $\mathrm{var}(\epsilon) = \sigma^2$. Independent runs are carried out, with m runs at each of $(-1, -1)$, $(-1, 1)$, $(1, -1)$ and $(1, 1)$. Find the least squares estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ and find $\mathrm{cov}\hat{\boldsymbol{\beta}}$. Use the General Equivalence Theorem to determine whether or not this design is D-optimal.

The above design is to be augmented by adding m further observations at a design point (x_1, x_2) with x_1 and x_2 in $\{-1, 0, 1\}$. The new design point is to be chosen such that the augmented design will be D-optimal among all such augmented designs. Let $\widetilde{X}$ be the X-matrix for the new design. By finding the determinant of $\widetilde{X}^T\widetilde{X}$, show how to determine suitable choices for x_1, x_2.