University of Strathclyde

53.483 DATA ANALYSIS I

Tuesday 22nd January 2008
10.00a.m. - 12noon

All questions may be attempted.
Credit will be given for the best THREE answers only.

1.(a) The mathematical model for the response variable X_{ij} in a completely randomised design with k treatments and n replications per treatment can be described by

$$X_{ij} = \mu + \tau_j + \varepsilon_{ij}, \qquad 1 \le i \le n, \;\; 1 \le j \le k,$$

where μ is the grand mean, τ_j the effect of treatment j and ε_{ij} the experimental error. Assume that treatments are fixed effects, $\Sigma\tau_j = 0$. Assume also that ϵ_{ij} are independent and follow a $N(0, \sigma_e^2)$ distribution. Give the definitions for SS_{treat} and MS_{treat}. Show that

$$E(MS_{treat}) = \sigma_e^2 + n\sigma_T^2,$$

where $\sigma_T^2 = (\tau_1^2 + \cdots + \tau_k^2)/(k-1)$. (You may use the following results without proof. Let Y_i $(1 \le i \le m)$ be independent and normally distributed $N(0, \sigma^2)$ variables. Define $\bar{Y} = (Y_1 + \cdots + Y_m)/m$. Then $\bar{Y} \sim N(0, \sigma^2/m)$ and

$$E\Big[\frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \bar{Y})^2\Big] = \sigma^2. \text{)}$$

(13 marks)

(b) The partially completed ANOVA table for a $p \times q$ design is shown below:

Source	SS	d.f.	MS	F
A	3.06		1.02	
B	159.50	3		
$A \times B$	11.79			
Error				
Total	180.44	31		

Assume A is a fixed effect but B is a random effect so this is a mixed model. Using a suitable notation, give a clear description of the model, including any assumptions necessary to undertake the analysis of variance. Fill in the blanks in the ANOVA table. Carry out the F-tests for interaction $A \times B$ and main effects A and B. Test using $\alpha = 0.05$.

(12 marks)

2. The following data were generated from a randomised block design with 4 treatments and 4 blocks:

	Block 1	Block 2	Block 3	Block 4
Treat 1	1.25	1.43	1.32	1.31
Treat 2	2.05	1.56	1.68	1.69
Treat 3	1.95	2.00	1.83	1.81
Treat 4	1.75	1.93	1.70	1.59

Assume that treatments and blocks are fixed effects. The classical mathematical model for the design is

$$X_{ij} = \mu + \beta_i + \tau_j + \varepsilon_{ij}, \quad 1 \le j \le 4, \ 1 \le i \le 4,$$

where μ is the grand mean, τ_j is the effect of treatment j, β_i is the effect of block i while ε_{ij} is the experimental error. The GLIM mathematical model for the design is

$$X_{ij} = \tilde{\mu} + \tilde{\beta}_i + \tilde{\tau}_j + \varepsilon_{ij}, \quad 1 \le j \le 4, \ 1 \le i \le 4.$$

(a) State the classical assumptions and the GLIM assumptions. State the relationships between the classical parameters and the GLIM parameters by expressing $\tilde{\mu}$, $\tilde{\beta}_i$ and $\tilde{\tau}_j$ in terms of μ, β_i and τ_j.

(3 marks)

(b) State the GLIM commands you would use to input the data in order to produce the following results:

(3 marks)

```
[i] ? $fit$
[o]      deviance - 0.95584
[o] residual df = 15
[i] ? $fit +block$
[o]      deviance = 0.89167 (change = -0.06417)
[o] residual df = 9         (change = -3)
[o]
[i] ? $fit + treat$
[o]      deviance = 0.17296 (change = -0.7187)
[o] residual df = 9         (change = -3)
[o]
[i] ? $ disp e s v r$
```

```
[o]      estimate      s.e.   parameter
[o] 1       1.399   0.09169           1
[o] 2   -0.02000   0.09802    BLOCK(2)
[o] 3    -0.1175   0.09802    BLOCK(3)
[o] 4    -0.1500   0.09802    BLOCK(4)
[o] 5    -0.4175   0.09802    TREAT(2)
[o] 6     0.5700   0.09802    TREAT(3)
[o] 7     0.4150   0.09802    TREAT(4)
[o] scale parameter 0.01922
[i] ? $tab the yield mean for treat$
[o] TREAT       1       2       3
[o] MEAN    1.327   1.745   1.898   1.742
```

(c) Based on the GLIM output, draw up the ANOVA table and state what conclusions can be drawn from it (tests using $\alpha = 0.05$).

(5 marks)

(d) Carry out the Newman-Keuls MRT for the 4 treatments (test using $\alpha = 0.05$) and state your conclusions.

(7 marks)

(e) Based on the GLIM output, the standard error (s.e.) for the GLIM parameter $\tilde{\tau}_2$ is 0.09802. State (without proof) the least-squares estimator for the $\tilde{\tau}_2$ in terms of the observations, and show its variance from first principles and hence verify its s.e. $= 0.09802$.

(7 marks)

3. Data are provided on content uniformity of film-coated tablets produced for a cardiovascular drug used to lower blood pressure. A random sample of 3 batches was chosen at each of two manufacturing sites, and from each batch 5 tablets were examined. We wish to know if there are differences between the two sites (fixed effects) and between batches within sites (random effects) i.e. mixed model.

	Site 1			Site 2		
	Batch 1	Batch 2	Batch 3	Batch 1	Batch 2	Batch 3
	5.03	4.64	5.10	5.05	5.46	4.90
	5.10	4.73	5.15	4.96	5.15	4.95
Repl	5.25	4.82	5.20	5.12	5.18	4.86
	4.98	4.95	5.08	5.12	5.18	4.86
	5.05	5.06	5.14	5.05	5.11	5.07

(a) State a linear model that may be used to represent the data and state the assumptions of the model.

(6 marks)

(b) State the GLIM commands you would use to input the data in order to produce the following results.

[i] ? fit

[o] deviance = 0.76247

[o] d.f. = 29

[o]

[i] ? $fit + site$

[o] deviance = 0.74421 (change = -0.01825)

[o] d.f. = 28 (change = -1)

[o]

[i] ? $fit + site.batch$

[o] deviance = 0.29020 (change = -0.4540)

[o] d.f. = 24 (change = -4)

(5 marks)

(c) Using the GLIM results provided, give the ANOVA table and undertake hypothesis tests to see if there are significant differences between the two sites and between batches within sites. Test using $\alpha = 0.05$.

(7 marks)

(d) Compute the 90% confidence interval for σ_e^2 (the variance of the experimental error).

(7 marks)

4.(a) In a 2^4 design with treatment A, B, C, D and 8 blocks and 3 replicates, the design is arranged so that AB, BC and CD are confounded with block effects in 3 replicates. Specify the units to go into each block and identify the composition of the ANOVA table.

(12 marks)

(b) In a 2^4-design, 3 replications are carried out for each of the following half of the 16 treatment combinations:

$$(1) \quad b \quad c \quad bc \quad ad \quad abd \quad acd \quad abcd$$

Identify the defining contrast and indicate aliases.

(9 marks)

(c) Identify the composition of the ANOVA table for part (b).

(4 marks)

Solutions to DA1 Exam 2007/8

1(a) $T_j = \sum_{i=1}^{n} X_{ij}$, $\quad \bar{T}_j = \frac{T_j}{n}$

$$G = \sum_{i=1}^{n} \sum_{j=1}^{n} X_{ij}, \quad \bar{G} = \frac{G}{nk}$$

$$SS_{treat} = n \sum_{j=1}^{k} (\bar{T}_j - \bar{G})^2$$

$$MS_{treat} = \frac{SS_{treat}}{k-1}$$

Note $\bar{T}_j = \frac{1}{n} \sum_{i=1}^{n} (\mu + \tau_j + \varepsilon_{ij})$

$$= \mu + \tau_j + \bar{e}_j$$

where $\bar{e}_j = \frac{1}{n} \sum_{i=1}^{n} \varepsilon_{ij} \sim N(0, \frac{\sigma_e^2}{n})$

$$\bar{G} = \frac{1}{nk} \sum_{i=1}^{n} \sum_{j=1}^{k} (\mu + \tau_j + \varepsilon_{ij})$$

$$= \mu + \frac{1}{nk} \sum_{i=1}^{n} \sum_{j=1}^{k} \varepsilon_{ij}$$

$$= \mu + \bar{e}$$

where $\bar{e} = \frac{1}{k} \sum_{j=1}^{k} \left(\frac{1}{n} \sum_{i=1}^{n} \varepsilon_{ij} \right) = \frac{1}{k} \sum_{j=1}^{k} \bar{e}_j$

$\sim N(0, \sigma_e^2/nk)$

Compute

$$E(MS_{treat}) = \frac{n}{k-1} E\left[\sum_{j=1}^{k} (\mu + \tau_j + \bar{e}_j - \mu - \bar{e})^2 \right]$$

$$= \frac{n}{k-1} E\left[\sum_{j=1}^{k} (\tau_j^2 + 2\tau_j(e_j - \bar{e}) + (e_j - \bar{e})^2 \right]$$

$$= \frac{n}{k-1}\left[\sum_{j=1}^{k} \tau_j^2 + \sum_{j=1}^{k} 2\tau_j (E\bar{e}_j - E\bar{\bar{e}}) + E\sum_{j=1}^{k}(\bar{e}_j - \bar{\bar{e}})^2\right]$$

$$= n\,\frac{\tau_1^2 + \cdots + \tau_k^2}{k-1} + 0 + n\,E\left[\frac{1}{k-1}\sum_{j=1}^{k}(\bar{e}_j - \bar{\bar{e}})^2\right]$$

$$= n\,\sigma_T^2 + n\cdot\frac{\sigma_e^2}{n} = n\,\sigma_T^2 + \sigma_e^2$$

1(b) This is a $p \times q$ design with $p = 4$ (A) and $q = 4$ (B)

$$X_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijk}$$

$$1 \le i \le 4, \quad 1 \le j \le 4, \quad 1 \le k \le 2$$

where μ – grand mean

α_i – effect of factor A level i

β_j – - - - - - - B - - - j

$\alpha\beta_{ij}$ – effect of interaction between factor A level i & factor B level j

ε_{ijk} – experimental error

$\because$ A is a fix effect while B is a random effect

$\therefore$ This a mixed-effects model

Assumptions:

α_i's constants, $\sum_{i=1}^{4} \alpha_i = 0$

$$\left.\begin{array}{l} \beta_j \sim N(0, \sigma_B^2) \\ \alpha\beta_{ij} \sim N(0, \sigma_{AB}^2) \\ \varepsilon_{ijk} \sim N(0, \sigma_e^2) \end{array}\right\} \text{independently}$$

ANOVA

Source	SS	d.f.	MS	F
A	3.06	3	1.02	$\frac{1.02}{1.31} = 0.77$
B	159.50	3	53.17	$\frac{53.17}{0.38} = 139.92$
AB	11.79	9 (=3×3)	1.31	$\frac{1.31}{0.38} = 3.44$
Error	6.09	16	0.38	
Total	180.44	31		

Hypothesis tests

AB. $H_0: \sigma_{AB}^2 = 0$

$H_a: \sigma_{AB}^2 > 0$

$$F = \frac{MS_{AB}}{MS_{error}} = \frac{1.31}{0.38} = 3.44$$

$F_{0.05}(9, 16) = 2.55 < F$

∴ reject H_0, i.e. AB is significant

A. $H_0: \alpha_i = 0$

$H_a: \alpha_i$'s differ

$$F = \frac{MS_A}{MS_{AB}} = \frac{1.02}{1.31} = 0.77$$

$F_{0.05}(3, 9) = 3.86 > 0.77$

∴ don't reject H_0 i.e. A is not sign.

B. $H_0: \sigma_B^2 = 0$

$H_a: \sigma_B^2 > 0$

$$F = \frac{MS_B}{MS_{error}} = \frac{53.17}{0.38} = 139.92$$

$F_{0.05}(3, 16) = 3.24 < 139.92$

∴ reject H_0. i.e. B is significant

2 (a) The classical assumptions:

μ, β_i, τ_j are all constants,

$$\sum_{i=1}^{4} \beta_i = 0 \ , \quad \sum_{j=1}^{4} \tau_j = 0$$

$\varepsilon_{ij} \sim N(0, \sigma_e^2)$ independently

The GLIM assumptions:

$\tilde{\mu}$, $\tilde{\beta}_i$, $\tilde{\tau}_j$ are all constants,

$$\tilde{\beta}_1 = 0 \ , \quad \tilde{\tau}_1 = 0$$

$$\tilde{\mu} = \mu + \beta_1 + \tau_1$$

$$\tilde{\beta}_i = \beta_i - \beta_1 \qquad (2 \le i \le 4)$$

$$\tilde{\tau}_j = \tau_j - \tau_1 \qquad (2 \le j \le 4)$$

(b)

```
[i] ? $uni 16$
[i] ? $data yield$
i] ? $read
[i] $REA? 1.25 1.43 1.32 1.31
[i] $REA? 2.05 1.56 1.68 1.69
[i] $REA? 1.95 2.00 1.83 1.81
[i] $REA? 1.75 1.93 1.70 1.59
[i] ? $ca block=%gl(4,1)$
[i] ? $ca treat=%gl(4,4)$
[i] ? $factor block 4 treat 4$
[i] ? $look yield block treat$
[o]        YIELD     BLOCK     TREAT
[o]    1   1.250     1.000     1.000
[o]    2   1.430     2.000     1.000
[o]    3   1.320     3.000     1.000
[o]    4   1.310     4.000     1.000
[o]    5   2.050     1.000     2.000
[o]    6   1.560     2.000     2.000
[o]    7   1.680     3.000     2.000
[o]    8   1.690     4.000     2.000
[o]    9   1.950     1.000     3.000
[o]   10   2.000     2.000     3.000
[o]   11   1.830     3.000     3.000
[o]   12   1.810     4.000     3.000
[o]   13   1.750     1.000     4.000
[o]   14   1.930     2.000     4.000
[o]   15   1.700     3.000     4.000
[o]   16   1.590     4.000     4.000
[i] ? $yvar yield$
```

2(c) ANOVA

Source	SS	df	MS	F
Block	0.06417	3	0.02139	1.11290
Treat	0.7187	3	0.23957	12.4646
Error	0.17296	9	0.01922	
Total	0.95583	12		

Hypothesis test for treatments.

$H_0: \tau_1 = \tau_2 = \tau_3 = \tau_4 = 0$

$H_a: \tau_j$'s differ

Given $\alpha = 0.05$ $F_{3,9}(0.05) = 3.86$

$\because F = 12.4646 > F_{3,9}(0.05)$

$\therefore$ reject H_0 and accept H_1 i.e. there is a significant difference among treat. means.

Hypothesis test for blocks.

$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

$H_1: \beta_i$'s differ

$\because F = 1.1129 < F_{3,9}(0.05)$

$\therefore$ don't reject H_0

3(a) MRT.

Treat	1	4	2	3
Mean	1.327	1.742	1.745	1.898

MSerror = 0.01922, d.f. = 9, s.e. of mean = sqrt(0.01922/4) = 0.06932

P	2	3	4
Range	3.2	3.95	4.42

P	2	3	4
LS range	0.2218	0.2738	0.3064

Tests
3 vs 1: 1.898-1.327=0.571>0.3064 => significant
3 vs 4: 1.898-1.742=0.156<0.2738 => not significant
3 vs 2: 1.898-1.745=0.153<0.2218 => not significant
2 vs 1: 1.745-1.327=0.418>0.2738 => significant
2 vs 4: 1.745-1.742=0.003<0.2218 => not significant
4 vs 1: 1.742-1.327=0.415>0.2218 => significant

Conclusions: treat 1 differs from treatments 2, 3, 4 while treatments 2, 3, 4 do not differ

2(e) From 2(a), we know $\tilde{\tau}_2 = \tau_2 - \tau_1$, while the l.s. estimators for τ_2 and τ_1 are $\bar{T}_2$ and $\bar{T}_1$, respectively, so the l.s. estimator for $\tilde{\tau}_2$ is

$$\bar{T}_2 - \bar{T}_1$$

where $\bar{T}_2 = \frac{1}{4}\sum_{i=1}^{4} X_{i2}$, $\bar{T}_1 = \frac{1}{4}\sum_{i=1}^{4} X_{i1}$

Since X_{ij}'s are independent random variable with variance σ_e^2,

$$Var(\bar{T}_2 - \bar{T}_1) = Var(\bar{T}_2) + Var(\bar{T}_1)$$
$$= \frac{1}{16}\sum_{i=1}^{4} Var(X_{i2}) + \frac{1}{16}\sum_{i=1}^{4} Var(X_{i1})$$
$$= \frac{1}{16}\sum_{i=1}^{4} \sigma_e^2 + \frac{1}{16}\sum_{i=1}^{4} \sigma_e^2$$
$$= \frac{\sigma_e^2}{2}$$

The s.e. of the l.s. estimator for $\tilde{\tau}_2$ is then $\sigma_e^2/\sqrt{2}$ Estimate σ_e by $s_e = \sqrt{MS_{error}} = 0.1386$.

Then the s.e. of the l.s. estimator for $\tilde{\tau}_2$ is

$$\frac{0.1386}{\sqrt{2}} = 0.09802$$

which is the same as the GLM output

3 a) This is a nested designs, since different batches are used at each site. The mathematical model is:

$$X_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \varepsilon_{ijk}.$$

$$1 \le i \le 2, \quad 1 \le j \le 3, \quad 1 \le k \le 5$$

where μ — grand mean

α_i — effect of site i

$\beta_{j(i)}$ — effect of batch j nested in ~~treatment i~~ site i

ε_{ijk} — experimental error.

From the given conditions, we know that sites are fixed effects but batches are random effects. The corresponding assumptions are:

α_1 and α_2 are constants, $\alpha_1 + \alpha_2 = 0$

$$\left.\begin{array}{l} \beta_{j(i)} \sim N(0, \sigma_B^2) \\ \varepsilon_{ijk} \sim N(0, \sigma_e^2) \end{array}\right\} \text{independently.}$$

6

b)

```
o]
i] ? $units 30
i] ? $data content$
i] ? $read
i] $REA? 5.03 5.10 5.25 4.98 5.05
i] $REA? 4.64 4.73 4.82 4.95 5.06
i] $REA? 5.10 5.15 5.20 5.08 5.14
i] $REA? 5.05 4.96 5.12 5.12 5.05
i] $REA? 5.46 5.15 5.18 5.18 5.11
i] $REA? 4.90 4.95 4.86 4.86 5.07
i] ? $ca site=%gl(2,15) :batch=%gl(3,5)$
i] ? $fac site 2 batch 3$
i] ? $yvar content$
```

may use

5

$ ca site=%gl(2,15) : batch=%gl(6,5) $

$ fac site 2 batch 6 $

c) ANOVA

Source	SS	d.f.	ms	F
Sites	0.01825	1	0.01825	$\frac{0.01825}{0.11350} = 0.16$
Batches within sites	0.4540	4	0.11350	$\frac{0.11350}{0.01209} = 9.39$
Error	0.2902	24	0.01209	
Total	0.76247	29		

7

Tests for

Sites: $F_{0.05}(1, 4) = 7.71 > 0.16$

$\Rightarrow$ Insignificant

Batches within sites: $F_{0.05}(4, 24) = 2.75 < 9.39$

$\Rightarrow$ Significant.

d) Note from the ANOVA that

$$s_e^2 = MS_{error} = 0.01209$$

$$d.f. \text{ for error} = 24$$

$\therefore$ the 90% C.I. for σ_e^2 is

$$\frac{24\, s_e^2}{\chi^2_{24}(0.05)} < \sigma_e^2 < \frac{24\, s_e^2}{\chi^2_{24}(0.95)}$$

From the χ^2-table,

$$\chi^2_{24}(0.05) = 36.42$$

$$\chi^2_{24}(0.95) = 13.85$$

~~$\therefore$ The C.I. is~~

$$\frac{24 \times 0.01209}{36.42} < \sigma_e^2 < \frac{24 \times 0.01209}{13.85}$$

$$0.00796 < \sigma_e^2 < 0.0209$$

(7)

4(a) From the def. contrasts AB, BC, CD we derive the following interactions are also confounded.

$AB \times BC \pmod 2 = AC$

$AB \times CD \pmod 2 = ABCD$

$BC \times CD \pmod 2 = BD$

$AB \times BC \times CD \pmod 2 = AD$

For all 3 replicates, the 16 treat. comb'ns are assigned into 8 blocks as follows.

(1) a b ab c ac bc abc
d ad bd abd cd acd bcd abcd

$\Downarrow$ By AB

Gp 1: (1) ab c abc d abd cd abcd
Gp 2: a b ac bc ad bd acd bcd

$\Downarrow$ By BC

Gp 1: (1) abc d abcd
Gp 2: ab c abd cd
Gp 3: a bc ad bcd
Gp 4: b ac bd acd

$\Downarrow$ By CD

Block 1 : (1) abcd
2 : abc d
3 : ab cd
4 : c abd
5 : a bcd
6 : bc ad
7 : b acd
8 : ac bd

ANOVA

Source	d.f.	
A	1	
B	1	
C	1	
D	1	
ABC	1	
ABD	1	
ACD	1	T2
BCD	1	
Blocks	7	
Replicates	2	
Blocks × Repl	14	
Error	16	
Total	47	

4(b) the $\frac{1}{2}$-replicate is

(1) b c bc ad abd acd abcd

A × (the $\frac{1}{2}$-repl.) (mod. 2) yields

a ab ac abc d bd cd bcd

= the other $\frac{1}{2}$-repl.

so $A \in$ the def. contrast

B × (the $\frac{1}{2}$-repl.) (mod 2) yields

b (1) bc b abd ad abcd acd

= the same $\frac{1}{2}$-repl.

so $B \notin$ the def. contrast

C × (the $\frac{1}{2}$-repl.) (mod 2) yields c bc (1) b acd abcd ad abd

= the same $\frac{1}{2}$-repl.

So C $\notin$ the def. contrast

D × (the $\frac{1}{2}$-repl) (mod. 2) yields

d bd cd bcd a ab ac abc

= the other $\frac{1}{2}$-repl.

So D $\in$ the defining contrast

Hence the def. contrast is AD

Factor/interaction	Aliase
A	D
B	ABD
C	ACD
AB	BD
AC	CD
BC	ABCD
ABC	BCD

(9)

4(c) ANOVA

Source	d.f.
Repl.	2
A or D	1
B or ABD	1
C or ACD	1
AB or BD	1
AC or CD	1
BC or ABCD	1
ABC or BCD	1
error	14
Total	23

(4)