MATH273001

This question paper consists of 8 printed pages, each of which is identified by the reference **MATH2730**.

New Cambridge Elementary Statistical Tables are provided. Only approved basic scientific calculators may be used.

Examination for the Module MATH2730
(January 2004)

ANALYSIS OF EXPERIMENTAL DATA

Time allowed: **2 hours**

Attempt not more than FOUR questions.
All questions carry equal marks.

1. The random effects model for the one-way analysis of variance with equal numbers of observations on each treatment is given by

$$y_{ij} = \mu + A_i + \varepsilon_{ij} \qquad\qquad i = 1, \ldots, t; \;\; j = 1, \ldots, n,$$

where the $A_i \sim N(0, \sigma_A^2)$, the $\varepsilon_{ij} \sim N(0, \sigma^2)$ and all terms are independent.

What do the terms μ, A_i, and ε_{ij} represent?

Let

$$\bar{y}_{i\bullet} = \frac{1}{n}\sum_{j=1}^{n} y_{ij} \qquad \text{and} \qquad \bar{y}_{\bullet\bullet} = \frac{1}{nt}\sum_{i=1}^{t}\sum_{j=1}^{n} y_{ij}.$$

Write down without proof the analysis of variance identity which partitions the total sum of squares $\sum_{i=1}^{t}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{\bullet\bullet})^2$ into two components. Give your answer in terms of n, y_{ij}, $\bar{y}_{i\bullet}$, and $\bar{y}_{\bullet\bullet}$. What names are usually given to the three components of your answer?

The mean square for treatments is given in the usual notation by

$$MS_T = \frac{1}{t-1}\sum_{i=1}^{t} n(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2.$$

Given that $E(\bar{y}_{i\bullet}^2) = \mu^2 + \sigma_A^2 + \sigma^2/n$ and $E(\bar{y}_{\bullet\bullet}^2) = \mu^2 + \sigma_A^2/t + \sigma^2/N$, where $N = nt$, prove that $E(MS_T) = \sigma^2 + n\sigma_A^2$.

Write down, without proof, unbiased estimators of σ^2 and σ_A^2 in terms of mean squares.

In a certain mill yarn is produced on several machines which are all supposed to operate at the same speed. It is suspected that there may be differences between the machines. Three machines are chosen at random and their operating speeds at startup on successive days are recorded, giving the results tabulated below.

		Speed y_{ij} (metres per sec.)					Total $y_{i\bullet}$
	1	29.8	29.7	29.3	29.1	29.2	147.1
Machine	2	27.6	28.1	28.9	27.9	27.9	140.4
	3	28.3	27.0	28.9	27.2	29.0	140.4

$$\sum_{i=1}^{3}\sum_{j=1}^{5} y_{ij}^2 = 12217.41.$$

Justify the choice of the random effects model.

Construct the ANOVA table for these data and use it to decide whether there are significant differences between the operating speeds of different machines.

Estimate the model parameters μ, σ^2, and σ_A^2. What name is given to the parameters σ^2 and σ_A^2?

2. In a one-way fixed effects analysis of variance model

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \qquad\qquad i = 1, \ldots, t; \;\; j = 1, \ldots, n,$$

let y_{ij} denote the jth observed value for the ith treatment Further, let $\bar{y}_{i\bullet}$ denote the sample mean of the ith treatment and $\bar{y}_{\bullet\bullet}$ denote the overall mean. Prove the identity

$$\sum_{i=1}^{t}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{\bullet\bullet})^2 = n\sum_{i=1}^{t}(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2 + \sum_{i=1}^{t}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i\bullet})^2,$$

and explain what each term represents.

The data below resulted from measuring the density (in tonnes per cubic metre) of three different types of granite used in constructing building facades.

Granite type		
Biotite	Hornblende	Tourmaline
2.21	2.17	1.99
2.30	2.12	1.96
2.25	2.02	1.87
2.09	2.19	2.05
2.18	2.35	1.81
2.36	2.24	2.10

Summary statistics are

$$\sum_{i=1}^{3}\sum_{j=1}^{6} y_{ij}^2 = 81.7338, \qquad \sum_{i=1}^{3}\frac{y_{i\bullet}^2}{6} = 81.5681, \qquad \sum_{i=1}^{3}\sum_{j=1}^{6} y_{ij} = 38.26,$$

where $y_{i\bullet}$ denotes the total of the observations on granite type i.

Justify the use of the one-way fixed effects ANOVA model to analyse these data. State the appropriate hypothesis to investigate whether the mean density is the same for each type of granite. Test your hypothesis by constructing an ANOVA table.

How would you compute the residuals for the granite density data? (It is not necessary to compute the residuals themselves, simply indicate how you would do so.)

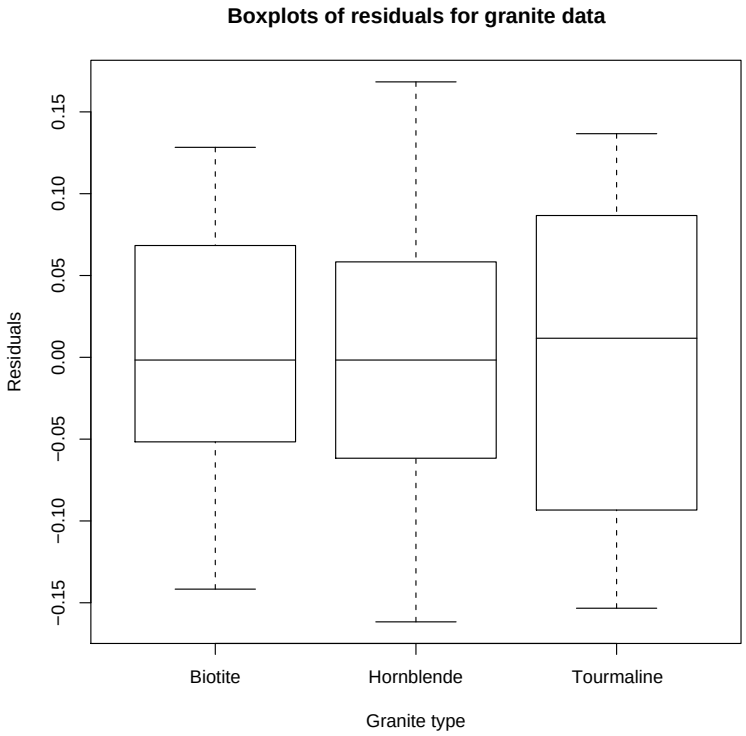
Boxplots of the residuals for each type of granite and a normal QQ plot are shown on the next page. What conclusions can you draw from these plots? For each plot, suggest an alternative that might convey the same information.

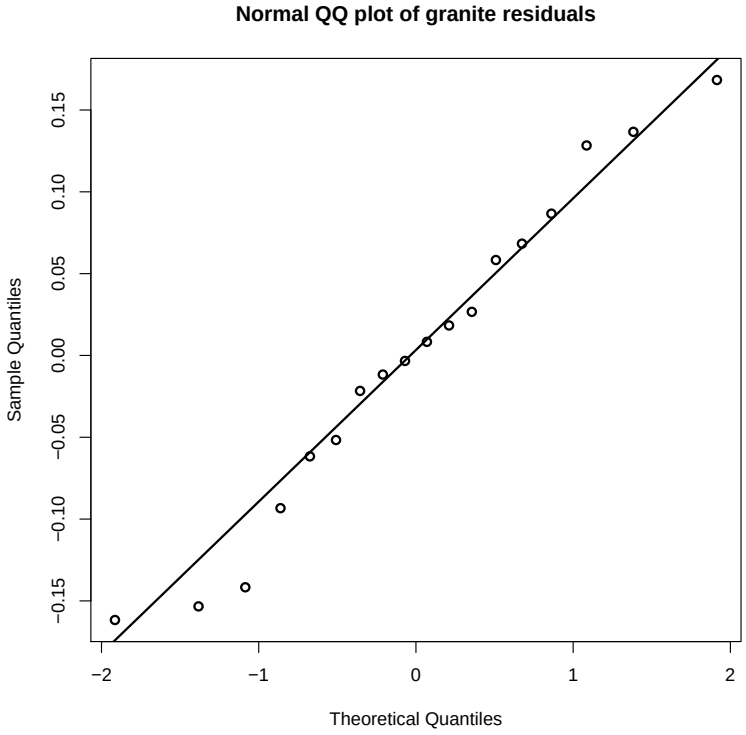
Given that the sample variances of the densities of the types of granite are, to 5 decimal places, $S_1^2 = 0.00894$, $S_2^2 = 0.01238$, and $S_3^2 = 0.01183$, use the following formulae to carry out Bartlett's test to determine whether the assumption of constant variance is valid. Under what circumstances would you not use Bartlett's test?

$$B = \frac{(N-t)\ln MS_E - \sum_i (n_i - 1)\ln S_i^2}{D}$$

where

$$D = 1 + \frac{\sum_i (n_i - 1)^{-1} - (N-t)^{-1}}{3(t-1)}.$$


Boxplots of residuals for granite data
Residuals
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
Biotite
Hornblende
Tourmaline
Granite type


Normal QQ plot of granite residuals
Sample Quantiles
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
−2
−1
0
1
2
Theoretical Quantiles

CONTINUED...

3. Epoxy-resin glue consists of an epoxy paste which is mixed with a hardening agent or resin to make it set. A construction company wishes to investigate the effect on setting time of five different resins, labelled A, B, C, D, and E. It is suspected that there may be differences between types of epoxy, so a randomised complete block design is used. The model is

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \qquad i = 1, \ldots, a; \;\; j = 1, \ldots, b,$$

where $\varepsilon_{ij} \sim N(0, \sigma^2)$ independently for all i and j and the parameters are constrained by the equations

$$\sum_{i=1}^{a} \alpha_i = 0 \qquad \text{and} \qquad \sum_{j=1}^{b} \beta_j = 0.$$

For the experiment, five batches of epoxy are used and each resin is applied to a portion of each batch of epoxy. The resulting setting times in minutes are given below.

		Epoxy batch				
		1	2	3	4	5
	A	8	11	7	6	10
	B	5	9	3	4	8
Resin	C	12	9	10	5	9
	D	4	4	1	1	6
	E	6	4	6	1	3

An analysis of variance using R produces the following ANOVA table where some values have been omitted.

```
> summary(aov(time ~ resin + epoxy, data = glue))
            Df  Sum Sq Mean Sq F value    Pr(>F)
resin        ? 133.040       ?       ? 0.0001966
epoxy        ?  57.440       ?
Residuals   16       ?   3.085
---
```

Fill in the missing values denoted by a "?" and interpret the results.

An experienced site foreman suspected even before the experiment that hardeners A and C took longer to set than hardeners D and E. Write down a sub-hypothesis in the form of a contrast which would test this, with the corresponding test statistic and its distribution under the null hypothesis.

Showing any calculations necessary, carry out the test corresponding to this sub-hypothesis and state your conclusion regarding the foreman's suspicions.

What is the maximum number of mutually orthogonal contrasts that could be constructed regarding the effects of resin on hardening time? Construct further contrasts to go with the one tested above to form a full set of mutually orthogonal contrasts and show that they are mutually orthogonal. You do not have to test these contrasts.

4. Consider the two-way fixed effects model given by

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \quad i = 1, \ldots, a; \;\; j = 1, \ldots, b; \;\; k = 1, \ldots, n.$$

with $\varepsilon_{ijk} \sim N(0, \sigma^2)$.

Constraints on the model parameters are needed to ensure that the parameters can be uniquely estimated. List a set of constraints that could be applied.

Show that

$$E\left(SS_B\right) = (b-1)\sigma^2 + an\sum_{j=1}^{b}\beta_j^2,$$

where SS_B denotes the sum of squares for factor B:

$$SS_B = an\sum_{j=1}^{b}(\bar{y}_{\bullet j \bullet} - \bar{y}_{\bullet\bullet\bullet})^2,$$

and where

$$\bar{y}_{i\bullet\bullet} = \frac{1}{bn}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk} \qquad \text{and} \qquad \bar{y}_{\bullet\bullet\bullet} = \frac{1}{abn}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk}.$$

In a study of wheat yields, three levels of each of two factors (fertiliser and insecticide) were selected. Fertiliser levels were "Low", "Medium" and "High", while brands "A", "B" and "C" of insecticide were used. A factorial experiment with two replicates was performed giving the following yields (in tonnes per hectare)

		Fertiliser		
		Low	Medium	High
Insecticide	A	7.64, 7.70	7.70, 7.64	7.79, 7.76
	B	7.55, 7.61	7.73, 7.79	7.73, 7.76
	C	7.70, 7.61	7.61, 7.67	7.82, 7.85

Why is the two-way layout model with fixed effects and interaction appropriate here?

These data were analysed in R with the following results:

```
> summary(aov(yield ~ insecticide * fertiliser, data = wheat))
                      Df  Sum Sq Mean Sq F value    Pr(>F)
insecticide            2 0.00070 0.00035  0.2188 0.8076719
fertiliser             2 0.06910 0.03455 21.5938 0.0003673
insecticide:fertiliser 4 0.03260 0.00815  5.0938 0.0201012
Residuals              9 0.01440 0.00160
---
```

Further calculation gave the following means:

Insecticide Brand	Mean
A	7.705
B	7.695
C	7.710

Fertiliser Dosage	Mean
Low	7.635
Medium	7.690
High	7.785

and the overall mean $\bar{y}_{\bullet\bullet\bullet} = 7.703$ (to 3 dp).

What conclusions can you draw from the analysis of variance table?

Estimate the overall mean μ and the main effect parameters $\alpha_i \ \ (i = 1, 2, 3)$ and $\beta_j \ \ (j = 1, 2, 3)$. Comment on your results.

How would you estimate the interaction parameters γ_{ij}? (It is not necessary to estimate these, merely to indicate how you would do so.)

5. Consider the regression model

$$y_i = \alpha + \beta(x_i - \bar{x}) + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{1}$$

where the ε_i have independent normal distributions with mean zero and variance σ^2. If $\underline{y} = (y_1, y_2, \ldots, y_n)^T$, $\underline{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$, $\underline{\theta} = (\alpha, \beta)^T$, and

$$X = \begin{pmatrix} 1 & x_1 - \bar{x} \\ 1 & x_2 - \bar{x} \\ \vdots & \vdots \\ 1 & x_n - \bar{x} \end{pmatrix},$$

the usual least squares estimator of $\underline{\theta}$ is $\widehat{\underline{\theta}} = (X^T X)^{-1} X^T \underline{y}$.

Write (1) in vector form and show that $\widehat{\underline{\theta}}$ is an unbiased estimator of $\underline{\theta}$.

Let V be the covariance matrix of $\underline{\theta}$, i.e. $V = E\left\{(\widehat{\underline{\theta}} - \underline{\theta})(\widehat{\underline{\theta}} - \underline{\theta})^T\right\}$. Show that

$$V = \begin{pmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{\sigma^2}{S_{xx}} \end{pmatrix},$$

where $S_{xx} = \sum_i (x_i - \bar{x})^2$. (You may use without proof the fact that $E(\underline{\varepsilon}\underline{\varepsilon}^T) = \sigma^2 I_n$, where I_n is the $n \times n$ identity matrix.)

From this, write down $var(\widehat{\alpha})$, $var(\widehat{\beta})$, and $cov(\widehat{\alpha}, \widehat{\beta})$.

The data below relate to measurements on spring (April/May) and autumn (September/October) rainfall in Leeds for each of ten consecutive years (measured in inches).

Spring rainfall (x)	1.6	5.3	2.8	9.6	6.7	1.5	5.4	8.5	4.1	3.9
Autumn rainfall (y)	4.6	6	2.9	11.1	8.2	1.3	9.1	10.2	5.2	8.3

For these data, find $\widehat{\alpha}$ and $\widehat{\beta}$. (You may assume that $\sum_i (x_i - \bar{x})^2 = 67.184$ and $\sum_i y_i(x_i - \bar{x}) = 69.774$).

If y_0 denotes the mean value of y when $x = x_0$ and we estimate y_0 by $\widehat{y}_0 = \widehat{\alpha} + \widehat{\beta}(x_0 - \bar{x})$, show that

$$var(\widehat{y}_0) = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right).$$

You should state, but not prove, any properties of $\widehat{\alpha}$ and $\widehat{\beta}$ that you use.

Given that $SS_{RES} = 20.465$, estimate σ^2 and determine a 95% confidence interval for the mean value of autumn rainfall, y_0, when spring rainfall is $x_0 = 7$.

If next year's spring rainfall was $x = 7$ inches, what would your prediction interval for next year's autumn rainfall be? Why does this differ from the confidence interval for y_0 given $x_0 = 7$?

END