THE ROYAL STATISTICAL SOCIETY

2004 EXAMINATIONS – SOLUTIONS

GRADUATE DIPLOMA

APPLIED STATISTICS

PAPER I

The Society provides these solutions to assist candidates preparing for the examinations in future years and for the information of any other persons using the examinations.

The solutions should NOT be seen as "model answers". Rather, they have been written out in considerable detail and are intended as learning aids.

Users of the solutions should always be aware that in many cases there are valid alternative methods. Also, in the many cases where discussion is called for, there may be other valid points that could be made.

While every care has been taken with the preparation of these solutions, the Society will not be responsible for any errors or omissions.

The Society will not enter into any correspondence in respect of these solutions.

Graduate Diploma, Applied Statistics, Paper I, 2004. Question 1

(i) In a (weakly) stationary time series, the first two moments of the distribution of any set of $X_{t_1}, X_{t_2}, ..., X_{t_k}$ is unchanged by a shift (δ) of the times t_1, t_2, … . (A strongly stationary time series has the whole distribution unchanged.) $E\left(X_t X_s\right)$ is a function of $|t-s|$ only. Thus $E\left(X_t\right)=\mu$ and $E\left[\left(X_t-\mu\right)\left(X_{t+k}-\mu\right)\right]=\gamma(k)$.

An autoregressive model, of order k, is given by

$$X_t=\phi_1 X_{t-1}+\phi_2 X_{t-2}+...+\phi_k X_{t-k}+\varepsilon_t$$

where ε_t is the random ("noise") term.

(ii) There appears to be no trend, but a cycle of period 11 (counting by the number of points plotted on the line). Variance looks roughly constant; "noise" is clearly present. Point 35 seems very low, out of line with the earlier general pattern – it may be an outlier.

(iii) The cyclical pattern of period 11 is evident, probably decaying waves. It is not stationary, so other patterns are not clear.

(iv) 11-point differences should remove the cyclic pattern – and they have done. Again no trend is evident, and there is approximately constant variance, noise or possibly a simple AR or MA model. The series looks weakly stationary. The point 35 is now even more pronounced as a low value, possibly an outlier.

(v) Check the value at 35. Could also check the robustness of the results to removing this value. If, on checking, it is found to be an error, replace the existing value by the correct one.

(vi) After taking the differences, cyclical patterns are much less pronounced. There are no spikes, except possibly at lag 2.

(vii) AR(1), AR(2), MA(1) and MA(2) could be tried, also "white noise" alone. Residuals from a well-fitting model should be uncorrelated, (independent) Normally distributed, with mean 0 and constant variance.

Graduate Diploma, Applied Statistics, Paper I, 2004. Question 2

(i) $E\left(\hat{\boldsymbol{\beta}}\right)=E\left[\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}\right]=\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T E(\mathbf{Y})=\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}=\boldsymbol{\beta}$.

(ii)

$$\begin{aligned}\operatorname{Var}\left(\hat{\boldsymbol{\beta}}\right)&=\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T.\operatorname{Var}(\mathbf{Y}).\left\{\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right\}^T\\&=\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T.\sigma^2\mathbf{I}.\left\{\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right\}^T=\sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\left\{\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right\}^T\\&=\sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}=\sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}.\end{aligned}$$

(Note: $\mathbf{X}$ is constant; $\left(\mathbf{X}^T\mathbf{X}\right)$ is symmetrical, and therefore so also is $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$.)

(iii) The left-hand side is $\Sigma Y_i^2-n\bar{Y}^2=\Sigma\left(Y_i-\bar{Y}\right)^2$, which is the total sum of squares corrected for the mean (or, the total sum of squares about the mean).

The first term on the right is the regression sum of squares, about the mean, or the amount of variation explained by the regression relationship.

The second term on the right is the residual (error) sum of squares. This is the amount of variation not explained by the regression (after estimating $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}}$).

The analysis of variance table is

Source of variation	Sum of squares	d.f.	Mean square
Regression of $\mathbf{Y}$ on $\mathbf{X}$	$\text{SS}_{\text{reg}} = \hat{\boldsymbol{\beta}}^T\mathbf{X}^T Y-n\bar{Y}^2$	p	SS_{reg}/p
Residual (error)	$\text{SS}_{\text{resid}} = \left(\mathbf{Y}-\hat{\mathbf{Y}}\right)^T\left(\mathbf{Y}-\hat{\mathbf{Y}}\right)$	$n-p-1$	$\text{SS}_{\text{resid}}/(n-p-1)$
Total	$\text{SS}_{\text{tot}} = \mathbf{Y}^T\mathbf{Y}-n\bar{Y}^2$	$n-1$	

The usual null hypothesis is $\beta_1 = \beta_2 = \ldots = \beta_p = 0$, i.e. that there is no regression relationship. The test statistic is

$$\frac{\text{SS}_{\text{reg}}/p}{\text{SS}_{\text{resid}}/(n-p-1)}.$$

With the assumptions stated in the question, the sampling distribution of this test statistic if the null hypothesis is true is $F_{p,\,n-p-1}$.

Solution continued on next page

Part (iv)

(a) $$\mathbf{H}^T = \left(\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right)^T = \left(\mathbf{X}^T\right)^T\left(\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right)^T\mathbf{X}^T = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T = \mathbf{H}.$$

$$\mathbf{HH} = \left(\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right)\left(\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right) = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$$

$$= \mathbf{H}.$$

(b) $$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y} = \mathbf{HY}.$$

(c) $\operatorname{Var}\left(\hat{\mathbf{Y}}\right) = \operatorname{Var}\left(\mathbf{HY}\right) = \mathbf{H}\operatorname{Var}\left(\mathbf{Y}\right)\mathbf{H}^T$ but Var(**Y**) = σ^2**I**, and **H** is symmetric

$= \sigma^2\mathbf{HH} = \sigma^2\mathbf{H}$.

Part (v)

(a) $$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{HY}$$

$= \left(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\right) - \mathbf{H}\left(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\right)$ but $\mathbf{HX}\boldsymbol{\beta} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta}$

$$= \boldsymbol{\varepsilon} - \mathbf{H}\boldsymbol{\varepsilon} = \left(\mathbf{I} - \mathbf{H}\right)\boldsymbol{\varepsilon}.$$

(b) $$E\left(\mathbf{e}\right) = E\left[\left(\mathbf{I} - \mathbf{H}\right)\boldsymbol{\varepsilon}\right] = \left(\mathbf{I} - \mathbf{H}\right)E\left(\boldsymbol{\varepsilon}\right) = \mathbf{0}.$$

(c) $$\operatorname{Var}\left(\mathbf{e}\right) = \operatorname{Var}\left[\left(\mathbf{I} - \mathbf{H}\right)\boldsymbol{\varepsilon}\right]$$

$= \left(\mathbf{I} - \mathbf{H}\right)\operatorname{Var}\left(\boldsymbol{\varepsilon}\right)\left(\mathbf{I} - \mathbf{H}\right)^T$ but Var($\boldsymbol{\varepsilon}$) = σ^2**I**, and **I** – **H** is symmetric

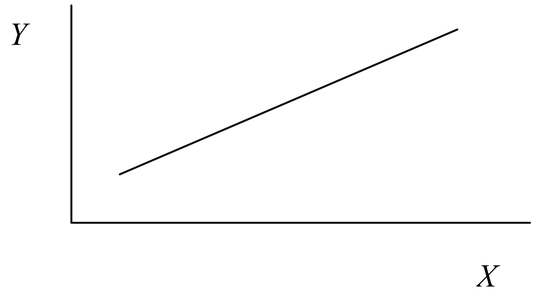
$$= \sigma^2\left(\mathbf{I} - \mathbf{H}\right)\left(\mathbf{I} - \mathbf{H}\right)$$

$$= \sigma^2\left(\mathbf{I}^2 - \mathbf{IH} - \mathbf{HI} + \mathbf{H}^2\right) = \sigma^2\left(\mathbf{I} - \mathbf{H} - \mathbf{H} + \mathbf{H}\right) = \sigma^2\left(\mathbf{I} - \mathbf{H}\right).$$

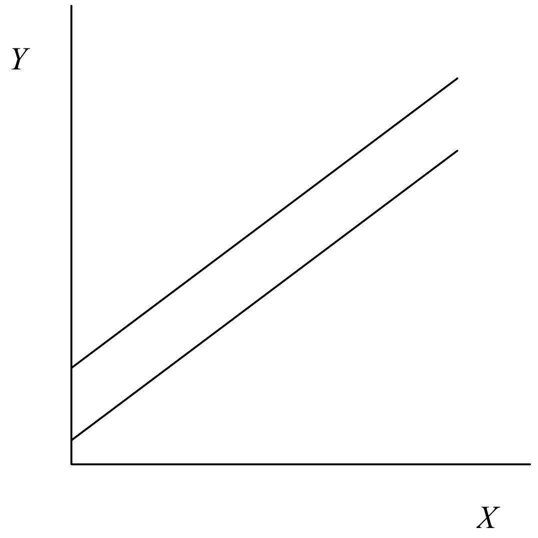
Graduate Diploma, Applied Statistics, Paper I, 2004. Question 3
(Solution continued on next page)

Part (i)

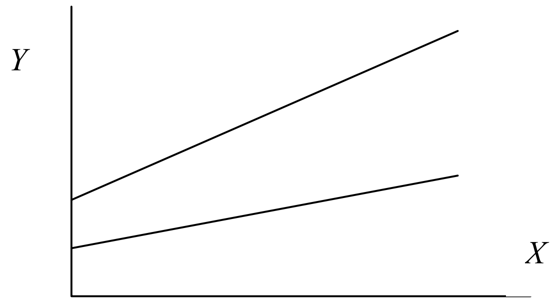
(a) is a single regression line, model $Y = a + bX$. (Could extend to a regression curve if thought appropriate.)



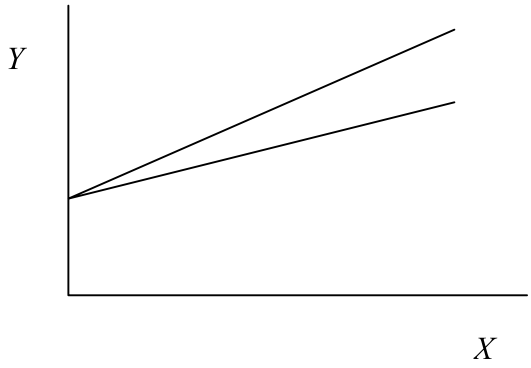
(b) has two parallel lines, one for each level of FAC, with different intercepts. Model $Y_j = a_j + bX_j$, $j = 1, 2$, for the two sets of points for the two levels of FAC.



(c) has two non-parallel lines – different slopes and different intercepts. Model $Y_j = a_j + b_jX_j$, $j = 1, 2$, for the two sets of points for the two levels of FAC.



(d) has two non-parallel lines with a common intercept. Model $Y_j = a + b_j X_j$, $j = 1, 2$, for the two sets of points for the two levels of FAC.



In all cases, a random term ε is added to the model, and all $\{\varepsilon_i\}$, $i = 1, 2, \ldots, n$ (for n observations) are assumed $N(0, \sigma^2)$, independent of one another.

Part (ii)

(a) This could be case (i)(b). There is a general linear relationship between performance and weight, with a negative slope. On average, performance is better for experienced drivers. For this set of data, more inexperienced drivers had heavier cars.

(b) Using forward selection, Models 1 and 2 are both superior to a model containing only a constant (a or a_j). Model 3 is better than either of these. However, Model 4 is not better. So use Model 3.

Assuming EXPER is a (0, 1) dummy variable as stated, the equations will be

$Y = 22.1 - 0.00181X$ for experienced drivers

$y = 21.5 - 0.00181X$ for inexperienced drivers,

where Y is KMPERL (performance) and X is WTKG (weight).

Summary: in each case, an increase of 1000 kg in weight leads to a decrease of 1.81 in km/l. On average, experienced drivers achieve 0.59 km/l better than inexperienced ones.

Graduate Diploma, Applied Statistics, Paper I, 2004. Question 4

(a) Tests and homework are likely to be different in character and it is unwise to assume that the teacher has data with an underlying multivariate Normal distribution. The response is binary (pass/fail). A logistic regression might be used, or a general linear model with binomial response and another link function. The data set is likely to be small, and last year's students may be different as a group from this year's. Obviously the same, or very similar, tests etc should be used, administered under similar conditions.

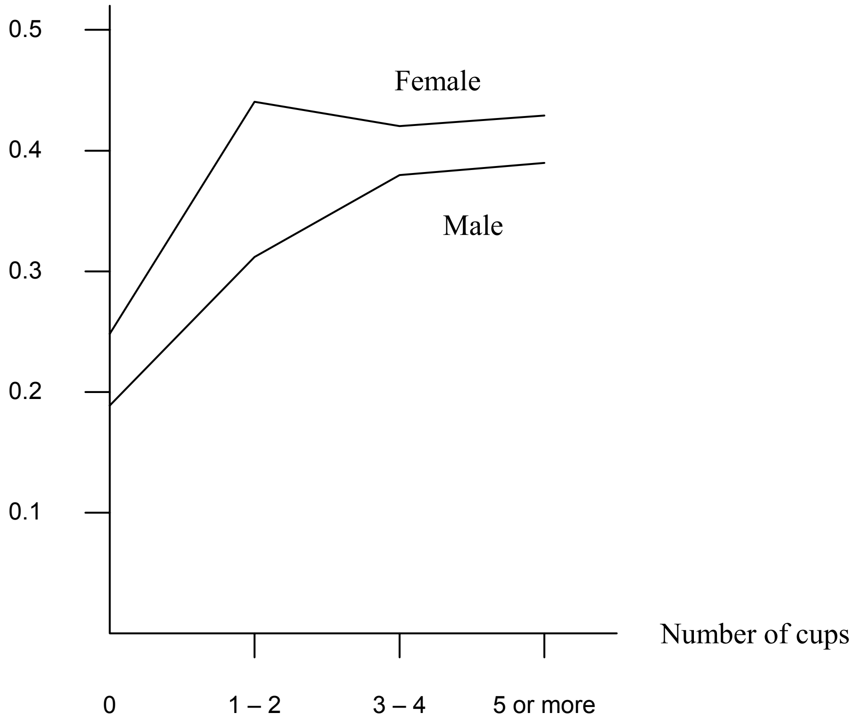
(b) This is a classification problem in which there are no prior groups. Likely variables that could determine purchasing habits have to be decided. It is hoped that enough of these were identified before any data were collected; if not, the results may be of limited use. A cluster analysis is indicated. The aim will be to distinguish between potential buyers of certain types of products, so as to target them when new items come into the catalogue.

(c) Multiple regression may be suitable. Covariates to separate any different treatments will be useful, as age-groups and patients at different levels of severity may react differently to the same treatment. Can a continuous response be assumed for quality of life? This may be satisfactory if a score based on several items is used, but often there may be doubt. Depending on the disease, there may be possible collinearity (e.g. between age and severity). The clinical relevance of all variables must be considered before they are included.

(d) This looks like a time series problem with a seasonal element. Conditions may have changed over the last 5 years, due perhaps to extra precautions being recommended to householders as a result of previous experience. Is the forecast purely of a number, or are different types of fire important? Is it important to know what equipment is used and what time is taken up dealing with a fire? A simple estimate of number of call-outs may not be adequate. The number of house fires per week is not likely to be very large, so how reliable will forecasts be, based on these data?

(i) The proportional incidence of cancer of the pancreas, classified by sex and number of cups, is as follows.

Number of cups	0	1 – 2	3 – 4	5 or more
Male	0.19	0.31	0.38	0.39
Female	0.25	0.44	0.42	0.43



The proportional incidence depends on sex. Initially it rises quickly with consumption but then stabilises. It stabilises more quickly and at a higher level for females than for males.

(ii) The data have been grouped, and are in any case difficult to describe by a continuous variable. The coding 0, 2, 4, 5 seems an odd choice, and is certainly not on a linear scale. Also, what does "5 or more" really mean?

A categorical variable would be better, if the raw data are not available.

Solution continued on next page

Part (iii)

(a) The general definition of the exponential family is that its probability density function or probability function, depending on a parameter θ, can be written as

$$f(x|\theta) = h(x)c(\theta)\exp\left(\sum_{i=1}^{k} w_i(\theta)t_i(x)\right)$$

where h, c, w, t are functions that vary from one member of the family to another.

For the binomial,

$$f(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$$

$$= \binom{n}{x}(1-\theta)^n\left(\frac{\theta}{1-\theta}\right)^x = \underset{\underset{h(x)}{\downarrow}}{\binom{n}{x}}\underset{\underset{c(\theta)}{\downarrow}}{(1-\theta)^n}\exp\left\{\underset{\underset{w(\theta)}{\downarrow}}{\log\left(\frac{\theta}{1-\theta}\right)}.\underset{\underset{t(x)}{\downarrow}}{x}\right\}$$

Here, we have $k = 1$ and $h(x)$ is only defined for $x = 0, 1, 2, \ldots, n$. Also, of course, $0 < \theta < 1$.

(b)

Variables in model	d.f.	Scaled deviance	Dev ÷ d.f.
–	7	27.373	3.91
*X*2	6	14.434	2.41
F1	4	8.969	2.24
F1 *X*2	3	2.051	0.68

For the model containing F1 and *X*2, deviance/d.f. is < 1, and this is by some way the best model.

(c) Given the reservations about the coding of *X*1, with the spurious suggestion of continuity given by use of (0, 2, 4, 5), the use of F1 is preferable as it is categorical. (A proper check of residuals should be made.)

(d) The odds ratio for females : males for a fixed rate of coffee consumption is $e^{0.35057} = 1.420$. The probability for females is significantly greater than the corresponding probability for males. An approximate 95% confidence interval is given by (e^a, e^b) where $a = 0.35057 - (1.96\times0.13363) = 0.0887$ and $b = 0.35057 + (1.96\times0.13363) = 0.6125$. Hence the interval is (1.093, 1.845).

Graduate Diploma, Applied Statistics, Paper I, 2004. Question 6

(i) Principal component analysis aims to describe the variation of a set of multivariate data in terms of a set of linear combinations of the original variables, called the principal components. The original data are (most likely) correlated but the principal components are not. When the original data are of different orders of magnitude, in different units, a few variables could dominate the whole calculation if carried out on the covariance matrix, but not if carried out on the correlation matrix.

(ii) If the original data $\mathbf{X}$ have correlation matrix $\mathbf{\Sigma}$, and the first principal component is $\mathbf{Z} = \boldsymbol{\beta}^T\mathbf{X}$ ($\boldsymbol{\beta}$ being a vector of coefficients), then $\text{Var}(\mathbf{Z}) = \boldsymbol{\beta}^T\mathbf{\Sigma}\boldsymbol{\beta}$.

The first principal component maximises $\boldsymbol{\beta}^T\mathbf{\Sigma}\boldsymbol{\beta}$ subject to $\boldsymbol{\beta}^T\boldsymbol{\beta} = 1$.

This is found by maximising $\boldsymbol{\beta}^T\mathbf{\Sigma}\boldsymbol{\beta} - \lambda(\boldsymbol{\beta}^T\boldsymbol{\beta} - 1)$. Differentiating and setting equal to zero gives $2\mathbf{\Sigma}\boldsymbol{\beta} - 2\lambda\boldsymbol{\beta} = \mathbf{0}$, or $(\mathbf{\Sigma} - \lambda\mathbf{I})\boldsymbol{\beta} = \mathbf{0}$. Thus $\boldsymbol{\beta}$ is an eigenvector of $\mathbf{\Sigma}$, corresponding to the eigenvalue λ.

(iii) (a) Times for adjacent sections of the race are highly positively correlated; correlations for other sections remain positive but are all smaller (around 0.5). The variances for the last two sections are high, and those for times 1 and 3 are higher than that for time 2.

(b) Since all the times are in the same units, the covariance matrix is a suitable base for the analysis. Relationships between the actual variables will be studied this way.

PC1 measures the total time to run the race, which gives greater weight to the last two stages. PC2 contrasts the last time with the earlier ones, especially 1 to 3. Together, these explain 85% of the variation.

(c) There is one high positive score. This implies that early times were short and later times long, i.e. the runner was faster at the start and slower at the end (which has positive weight in the score). Apart from this one case, there is a reasonably random pattern of scores against positions, and runners probably have different strategies.

(d) Roughly linear with a positive gradient, because it is related to the total time taken and therefore of course to the finishing position.

(e) One possible method is to classify into groups for ages and construct scatter plots of the PC scores with different symbols for age-groups. Another is to include age as an extra variable and carry out another principal component analysis, this time using the correlation matrix because of different units.

Graduate Diploma, Applied Statistics, Paper I, 2004. Question 7

Part (i)

Cluster analysis of multivariate data explores whether there are natural subgroups within the data, and if possible identifies what they are (in a somewhat subjective way).

Part (ii)

Methods for cluster analysis are not scale-independent. So a small number of variables with high variance can dominate a calculation of the distance matrix, and thus largely determine the clusters.

Part (iii)

(a) We have that $\bar{x}_1 = 1.395$ and Sd(x_1) = 0.448. $x_1(1)$ = 0.95 and so the corresponding standardised value is (0.95 – 1.395)/0.448 = –0.99 (to 2 d.p.).

Using the standardised data, the (Euclidean) distance between observations 1 and 2 is

$$\left[\sum_{i=1}^{5}\left(x_{1S}(i)-x_{2S}(i)\right)^2\right]^{1/2} = \left[(-0.99+0.95)^2+(-1.31+1.10)^2+\ldots+(0.02+1.52)^2\right]^{1/2}$$

$$= 1.783,$$

i.e. 1.79 approximately (the given figures will have been calculated using higher decimal accuracy than this).

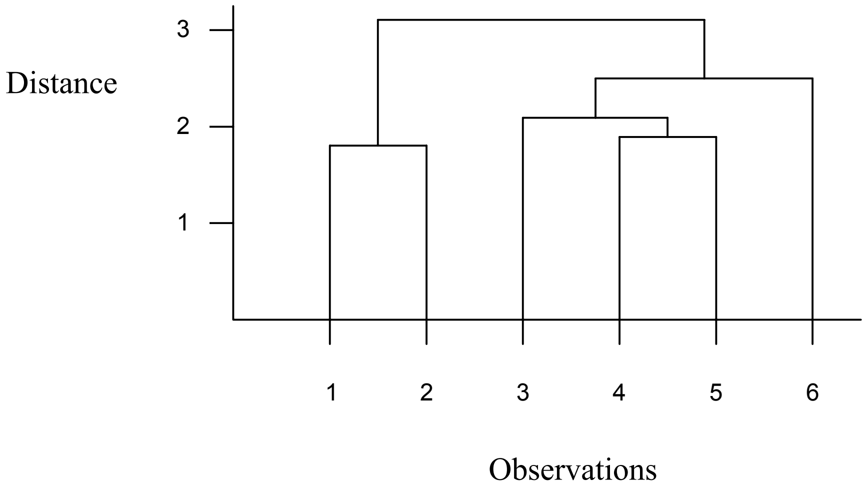
(b) Cluster 1 is observations (1, 2) at distance 1.79.

Cluster 2 is observations (4, 5) at distance 1.89.

Cluster 2 is joined by observation 3 at distance 2.08 and then by observation 6 at distance 2.50.

Finally clusters 1 and 2, i.e. (1, 2) and (6, 3, 4, 5), join at distance 3.02.

The dendrogram and the solution to part (iii)(c) are on the next page



(c) Figure (i) has a cluster formed of observations (3, 4, 5, 6), somewhat isolated from 1 and 2.

Figure (ii) has two clusters, (1, 2, 6) and (3, 4, 5).

Figure (iii) has two possible interpretations:

either two clusters, (1, 2) and (3, 6, 4, 5)

or three clusters (1, 2), (3, 6) and (4, 5).

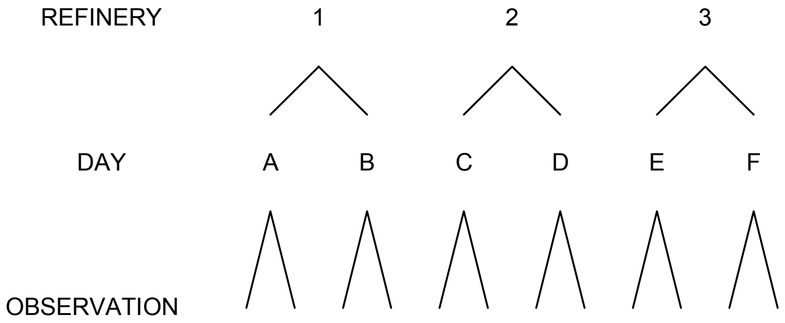
Single linkage uses minimum distance, and complete linkage uses maximum distance, between an observation in one cluster and an observation in another.

Observation 6 is a long way from observation 4 in the distance matrix but is close to observation 5. In single linkage (figure (i)), 6 goes with (3, 4, 5); but in complete linkage (on the standardised data, figure (ii)), it goes with (1, 2).

The raw data give a different distance matrix. For the raw data, figure (iii), complete linkage, is markedly different from figure (i), single linkage.

Graduate Diploma, Applied Statistics, Paper I, 2004. Question 8

(i) The data form a hierarchical set, as shown below.



One observation seems likely to be an outlier: ↑

Otherwise, there is some (small) difference between refineries and less between days or men.

(ii) $y_{ijk} = \mu + r_i + d_j + \varepsilon_{ijk}$ $\quad i = 1, 2, 3;\ \ j = 1, 2;\ \ k = 1, 2.$

In this model, y_{ijk} is a measured response; μ is the overall grand mean; r_i is the effect due to refinery i; d_j is the effect due to day j (some authors write this as $d_{(i)j}$ to emphasise that it refers to day j within the ith refinery); and ε_{ijk} is the residual (likewise sometimes written as $\varepsilon_{(ij)k}$).

The r_i and d_j are random variables with underlying distributions as follows:

$$r_i \sim \mathrm{N}\left(0, \sigma_R{}^2\right), \qquad d_j \sim \mathrm{N}\left(0, \sigma_D{}^2\right).$$

The residuals ε_{ijk} are also random variables, as usual:

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

The random variables in each set are mutually independent, and the sets are independent of each other.

The variance components $\left(\sigma_R{}^2, \sigma_D{}^2, \sigma^2\right)$ are to be estimated. F tests of hypotheses that any of these is zero can also be carried out.

Solution continued on next page

(iii) The refinery totals are 62, 47, 73. The day totals are 29, 33, 24, 23, 30, 43.

The grand total is 182, so the "correction factor" is $\frac{182^2}{12} = 2760.3333$.

Total sum of squares = $15^2 + 14^2 + \ldots + 19^2 - 2760.3333 = 149.6667$.

"Refineries" sum of squares = $\frac{62^2}{4} + \frac{47^2}{4} + \frac{73^2}{4} - 2760.3333 = 85.1667$.

Overall sum of squares between days = $\frac{29^2}{2} + \frac{33^2}{2} + \ldots + \frac{43^2}{2} - 2760.3333 = 131.6667$.

So the sum of squares "between days within refineries" is 131.6667 – 85.1667 = 46.5000.

Hence we get the analysis of variance below, in which the expected values of the mean squares are also shown.

Source of variation	d.f.	Sum of squares	Mean square	$E[MS]$
Between refineries	2	85.1667	42.583	$\sigma^2 + 2\sigma_D^2 + 4\sigma_R^2$
Between days within refineries	3	46.5000	15.500	$\sigma^2 + 2\sigma_D^2$
Between days	5	131.6667		
Between workmen within days (i.e. residual)	6	18.0000	3.000	σ^2
Total	11	149.6667		

The expected values of the mean squares show that

to test the null hypothesis $\sigma_D^2 = 0$, refer the mean square ratio 15.500/3.000 (= 5.17) to $F_{3,6}$. The upper 5% point is 4.76, so this null hypothesis is rejected

to test the null hypothesis $\sigma_R^2 = 0$, refer the mean square ratio 42.583/15.500 (= 2.75) to $F_{2,3}$. The upper 5% point is 9.55, so this null hypothesis is not rejected.

The estimates are $\hat{\sigma}^2 = 3.00$, $\hat{\sigma}_D^2 = 6.25$, $\hat{\sigma}_R^2 = 6.77$. These suggest a similar contribution to variability by days and by refineries, though the hypothesis tests have indicated that days are more variable than refineries. There are really too few degrees of freedom for reliable inference.

(iv) The observation 24 may be incorrect and, if possible, this should be checked. If it is incorrect (by being too large), the base level σ^2 is overestimated.

Treating such a small set of data as "continuous" must be open to doubt.