THE ROYAL STATISTICAL SOCIETY

2003 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, 2003. Question 1

(i) Series 1 : There is no pattern. All values in ACF and PACF lie in the range $\pm 2/\surd N = \pm 2/\surd 50 = \pm 0.28$. This suggests a random series of data, with "white noise" only. In the notation of part (ii),

$$X_t = \mu + Z_t .$$

Series 2 : The ACF gradually dies off (though increasing again from lag 10). The PACF has a spike at lag 1 and no other notable features, so an $AR(1)$ process is a suitable model:

$$X_t = \alpha X_{t-1} + Z_t \quad \text{(a Markov process),}$$

where α (of absolute value less than 1) is a constant.

Series 3 : The ACF begins with high positive values and these decrease only slowly. The PACF has a high value at lag 1 only. This process seems to consist of trend + white noise.

(ii) $E(X_t) = E(Z_t + 0.8Z_{t-1} - 0.4Z_{t-2}) = E(Z_t) + 0.8E(Z_{t-1}) - 0.4E(Z_{t-2}) = 0$, since $E(Z_t) = 0$ for all t.

Since the Z_t are independent ("white noise"),

$$\begin{aligned}\operatorname{Var}(X_t) &= \operatorname{Var}(Z_t) + (0.8)^2 \operatorname{Var}(Z_{t-1}) + (-0.4)^2 \operatorname{Var}(Z_{t-2}) \\ &= (1 + 0.64 + 0.16)\sigma_Z^2 = 1.8\sigma_Z^2 \ ;\end{aligned}$$

$$\begin{aligned}\operatorname{Cov}(X_t, X_{t-1}) &= \operatorname{Cov}(Z_t + 0.8Z_{t-1} - 0.4Z_{t-2} \ , \ Z_{t-1} + 0.8Z_{t-2} - 0.4Z_{t-3}) \\ &= 0.8\operatorname{Var}(Z_{t-1}) - 0.32\operatorname{Var}(Z_{t-2}) \quad \text{[by independence]} \\ &= 0.48\sigma_Z^2 = \gamma_X(1) \ ;\end{aligned}$$

$$\begin{aligned}\operatorname{Cov}(X_t, X_{t-2}) &= \operatorname{Cov}(Z_t + 0.8Z_{t-1} - 0.4Z_{t-2} \ , \ Z_{t-2} + 0.8Z_{t-3} - 0.4Z_{t-4}) \\ &= -0.4\operatorname{Var}(Z_{t-2}) \quad \text{[by independence]} \\ &= -0.4\sigma_Z^2 = \gamma_X(2) \ ;\end{aligned}$$

For $k \ge 3$, $\operatorname{Cov}(X_t, X_{t-k}) = 0$.

So

$$\gamma_X(k) = \begin{cases} 1.8\sigma_Z^2 & k = 0 \\ 0.48\sigma_Z^2 & k = 1 \\ -0.4\sigma_Z^2 & k = 2 \\ 0 & k \ge 3 \end{cases} \quad \text{and} \quad \rho_X(k) = \begin{cases} 1 & k = 0 \\ 0.27 & k = 1 \\ -0.22 & k = 2 \\ 0 & k \ge 3 \end{cases} .$$

Graduate Diploma, Applied Statistics, Paper I, 2003. Question 2

(a) Although the groups were selected at random, the diagram suggests that more of the "standard" group had longer times in the test before training; this may be important when carrying out the analysis, which needs to be able to correct for any possible practical bias.

The first t test looks only at the final times, and suggests that there is a very small difference between them, certainly not significant, but it takes no account of possible effects of the initial selection and so is not relevant.

Actual times have dropped substantially in each group, presumably through the effect of training (by either method) and through practice. The "new" group would have had less scope for change if their original times were indeed on the whole lower. The second t test does not take account of initial ability. It shows an estimated effect favouring the "standard" group, but this is likely to be an overestimate.

The analysis of covariance does take account of the information given by the original times. It shows a statistically significant advantage to the "standard" group, of 1.24 seconds on average, after correcting for initial times. This is statistically significant only at the 5% level (of the "usual" significance levels) but is probably the best indicator of the practical difference between the methods.

(b) Heights and weights of small children are very likely to be correlated, and therefore when one of them is used as the independent variable in a linear model the other will not add much if any further information. The third model is using both height and weight to predict catheter length. The first two models use only one of them.

All three models fit about equally well, as measured by R^2, the proportion of variability (among the 12 length measurements) that is explained. S is also about the same for each. Looking at the accuracy of the estimates of the coefficients in the regressions of l on h and w, we see that in the first two models the coefficients of h and w respectively have quite small standard errors, with the standard error of "constant" smaller in the second model (l on w). The third model, $l = a + b_1h + b_2w$, by contrast has relatively large standard errors for a, b_1 and b_2. Also the values of b_1 and b_2 are less easy to interpret; for example, b_1 is the average increase in l for unit increase in h when w is kept constant. The third model does not seem worth using, not giving any better predictions of l than the others.

The model $l = a + bw$ might be slightly better than $l = a' + b'h$ because the coefficients are better estimated (smaller standard errors). In any case, w might be easier to measure accurately for small children than h is. If so, this model is also preferred on practical grounds – and these should certainly be considered as well as statistical information.

Graduate Diploma, Applied Statistics, Paper I, 2003. Question 3

(i) The alternative to MANOVA would be separate analyses of each response, X_1, X_2 and X_3. But there may be effects or interactions that are significant when X_1, X_2, X_3 are considered together that would not be when looked at separately. Unnecessary multiple testing is also avoided.

It must be assumed that (X_1, X_2, X_3) have a multivariate Normal distribution, with equal variance-covariance structure in each group of data. Normality of responses within each X_i is necessary but not sufficient for this, and it is very difficult to check multivariate Normality, especially in small groups as in this example. The covariance matrices for each group could be compared, and any background or previous knowledge about the responses would be useful.

(ii) Usually MANOVA F tests are only approximate, and the various criteria have different characteristics in respect of robustness to departures from assumptions and of power. Although Wilks' Λ is often preferred, agreement among several criteria gives greater confidence in the answer.

(iii)

$$\Lambda = \frac{|\text{SSCP matrix for Error (Residual)}|}{|\text{SSCP matrix for Error } + \text{ SSCP matrix for Extrusion}|}$$

$$= \begin{vmatrix} 1.764 & 0.020 & -3.070 \\ 0.020 & 2.628 & -0.552 \\ -3.070 & -0.552 & 64.924 \end{vmatrix} \Bigg/ \begin{vmatrix} 1.764+1.740 & \ldots & \ldots \\ 0.020-1.504 & \ldots & \ldots \\ -3.070+0.855 & \ldots & \ldots \end{vmatrix}$$

(iv) Using the note on the F distribution associated with Λ, we have p = 3 and r = 16 (19 d.f. from 20 observations, less 3 for **extr**, **addit** and **extr*addit**), so $F_{3,14}$ is to be used. (5% point is 3.34, 1% point is 5.56.)

H_0 : interaction is zero $\qquad H_1$: interaction is non-zero

value of test statistic is 1.339, not significant, H_0 not rejected.

H_0 : **extr** effect zero $\qquad H_1$: **extr** effect not zero

value of test statistic is 7.554, significant at 1%, H_0 rejected.

H_0 : **addit** effect zero $\qquad H_1$: **addit** effect not zero

value of test statistic is 4.256, significant at 5%, H_0 rejected.

Hence the change in rate of extrusion and the amount of additive both affect the responses, but they do so independently without interaction.

(v) The univariate analyses could be carried out, and their results studied. Perhaps simultaneous confidence intervals could be calculated, or canonical variate analysis attempted. The absence of interaction makes it possible to study main effects in useful, standard ways.

Graduate Diploma, Applied Statistics, Paper I, 2003. Question 4

(i) The original variables, the answers to the questions, are likely to be highly correlated. Principal component analysis (PCA) gives linear combinations of the variables that are uncorrelated. The first PC accounts for the largest amount of variation in the data, the second for the next largest, and so on. If the questions form themselves into relatively distinct clusters then PCs are useful to define subsets, and possibly to suggest ways of combining scores.

PCs are only strictly valid for numeric data, but the data here are nearer to being categorical – at best ordinal. However, PCA is often used for data such as these.

(ii) A cluster analysis could be useful, using correlations (or absolute values of them); perhaps indications of the grouping of questions would be given.

(iii) PCA only works on complete records. If a respondent's answer to one question is missing, that whole set of responses will be omitted. Because PCA is based on analysis of variability in data, missing values cannot easily be imputed. The choice in this case is between analysing a large number of responses on a small number of questions and a small number of responses on a large number of questions. The strategy proposed seems sensible.

(iv) The first three eigenvalues add to 5.14, i.e. 5.14/6 or 85.7% of the total variation, and should be enough.

The first PC (54% of total variation) is an overall score of concern about cost – note that the "direction" of questions 2, 3, 4 is opposite to that of 1, 5, 6. The second PC (23% of total variation) measures the tendency of respondents to answer all questions in the same way, i.e. with similar scores. The third PC (9% of total variation and so relatively much less important) is dominated by question 4, perhaps contrasting its answers with those for question 2, perhaps also taking question 5 into account. The first two PCs therefore give most of the useful, easily understood, information.

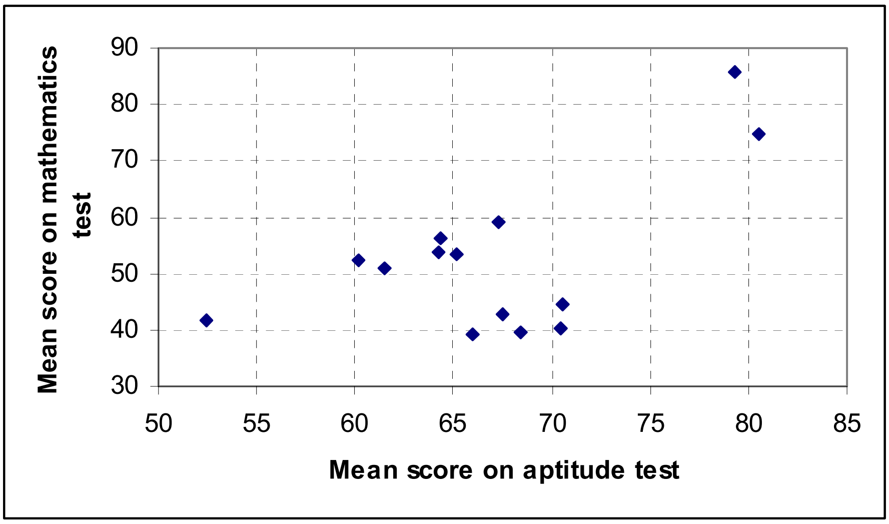
(v) The two unsatisfactory features of the data are the large amount of missing information, leading to 9 of the 15 questions being discarded, and the suggestion from the second PC that the respondents do not complete the form validly. Hence these results are not reliable. A fresh start is needed, with reworded questions and boxes to tick as in a survey.

Graduate Diploma, Applied Statistics, Paper I, 2003. Question 5

(i)(a) The least squares estimators of the coefficients of a linear model have minimum variance among all linear unbiased estimators. (This assumes that the residual error terms are uncorrelated random variables with common variance.)

(b) Weighted least squares should be used when the errors are uncorrelated but have different variances, e.g. if the variance is some function of the mean.

(ii)(a)



Note. False "origin" is placed at (50, 30).

It might be useful to add (perhaps in brackets alongside each point) the number of pupils entered by the school.

From the scatter plot, note that the two schools with the smallest number of pupils (these are the schools with aptitude test values near 80) have scores much higher than the others. Otherwise the scatter seems fairly random. However, the two highest points will be very influential in fitting a regression. (A cluster analysis here might perhaps suggest three clusters – low maths, middle maths and high maths.)

(b) Simple linear regression gives a poor fit ($R^2 = 38\%$). The value of "constant" is very poorly determined, though $p = 0.019$ for the coefficient of "aptitude" seems to suggest some linear relationship. (Fuller output, with information on influence and leverage, would be useful.)

However, the weighted regression, using numbers of pupils as weights, is even less satisfactory, giving an even lower R^2 and no evidence of a linear relationship ($p = 0.193$).

The mean maths score has variance σ_i^2/n_i where σ_i^2 is the within-school variance. The weighting assumes that all the σ_i^2 are similar, because then n_i is a suitable weight. But the σ_i^2 are not likely to be (approximately) equal, and until we have all the individual marks we cannot obtain the alternative weighting factors n_i/σ_i^2. (For example, the schools with small n_i might have selected pupils, leading to smaller σ_i^2 than the others.)

Neither regression is adequate, although the unweighted one might be a fair reflection of what is seen from the graph. Without more "diagnostic" information, we cannot go any further.

Graduate Diploma, Applied Statistics, Paper I, 2003. Question 6

(i) There will be a response (dependent) variable Y and a set $\boldsymbol{x}$ of possible explanatory (independent) variables, some or all of which can help to explain Y. The resulting model (apart form the "error" term) will be $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$ if p of the possible members of $\boldsymbol{x}$ are used.

Begin with fitting $Y = \beta_0$. Then fit in turn $Y = \beta_0 + \beta_i x_i$ for each i, where x_i is one of the set $\boldsymbol{x}$. If none of these shows a β_i which is significantly different from 0, there is no model better than "Y = mean + random error". Otherwise, choose that x_i which reduces the variation as much as possible (gives the smallest error (residual) sum of squares, or equivalently gives the greatest R^2). Call this x_1.

Next examine every possible two-variable regression including x_1, i.e. $Y = \beta_0 + \beta_1 x_1 + \beta_s x_s$ where x_s is any member of $\boldsymbol{x}$ other than x_1. On the basis of the extra sum of squares accounted for by x_s, choose the best x_s to include in the model; or, if no x_s gives a significant reduction in the residual sum of squares, stop at the one-variable model.

Continue in this way fitting extra terms as long as an $\boldsymbol{x}$-variable can be found that gives a significant reduction in the residual sum of squares compared with the existing model.

A good model selection procedure should provide as good an explanation of Y as possible using as few $\boldsymbol{x}$-variables as possible. This model will be easiest to apply and interpret. The drawback of the forward selection procedure is that once a particular $\boldsymbol{x}$-variable is in the model it cannot be removed; an optimal model may then not be reached, because there could be a pair (or perhaps a larger set) of $\boldsymbol{x}$-variables which together would be better even though neither gets into the model by itself. Thus a variable already in the model may be retained to the exclusion of other variables that would have been more useful.

[Putting this another way, suppose x_1 is the first variable to enter the model, so that x_1 gives the best one-variable model. Forward selection will now never select models that do not include x_1. However, there may be a pair (or a larger set) of other variables that would have given a better model than either x_1 alone or any other model that includes x_1.]

(ii)(a) Clearly X_2 enters first, because it makes the largest reduction in the error sum of squares. Once it is there, X_3 is better than X_1 to add to it in the model.

Step 1 (entering X_2) leaves an error SS of 117.17 with 22 d.f., thus the error MS is 5.326. So the 1 d.f. reduction here is 170.85 – 117.17 = 53.68. Thus we have an "extra sum of squares" test statistic of 53.68/5.326 = 10.08 which on comparing with $F_{1,22}$ is significant at 1%. So X_2 is retained in the model.

Now adding X_3 gives a further reduction of 117.17 – 90.007 = 27.163, and the error MS is 90.007/21 = 4.286. The $F_{1,21}$ test statistic is 27.163/4.286 = 6.34 which is significant at 5%. So X_3 is also retained in the model.

Adding X_1 to this two-variable model would reduce the error SS by only 90.007 – 88.453 = 1.554. This is less than the 20 d.f. error MS of 88.453/20 = 4.423. So we do not add X_1; we stop at X_2 and X_3.

Thus the model is $Y = \beta_0 + \beta_2 x_2 + \beta_3 x_3$.

The null hypothesis at each stage is that the sum of squares removed is not greater than that which remains as error mean square. The (one-sided) alternative hypothesis is that it is greater.

See next page for solution to (ii)(b)

(ii)(b) The calculations of the C_p statistic for each model are as follows. The quantity 4.4227 is the error mean square from the full model.

Model	s	$n-2s$	$SS_E/4.4227$	$C_p(s)$	
(1)	1	22	38.6302	16.63	
X_1	2	20	37.5267	17.53	
X_2	2	20	26.4929	6.49	
X_3	2	20	27.6822	7.68	
X_1, X_2	3	18	26.2962	8.30	
X_1, X_3	3	18	27.5284	9.53	
X_2, X_3	3	18	20.3511	2.35	← forward selection model
X_1, X_2, X_3	4	16	20	4.00	

A good model has $C_p(s) \approx s$ (which has of course to be true for the full model from which the 4.4227 was calculated). Clearly the forward selection model is best on this criterion, and the full model contributes very little to the explanation of Y that is not already contained in (X_2, X_3).

Graduate Diploma, Applied Statistics, Paper I, 2003. Question 7

(i) Linear models assume that the residual (error) term included in a model is a random variable having constant variance for all values of the response variable Y. Sometimes a response Y is known not to have constant variance, and sometimes there is a relation between expected value and variance which is known or which can be deduced from a plot of the residuals. As shown in part (ii), a function $f(y)$ can often be found from this relation such that Var($f(y)$) is constant. Analysis is then carried out in terms of $f(y)$, not y; f is a transformation to stabilise variance.

For example, if variability is proportional to the size of response, a log transformation will often stabilise the variance, i.e. Var(log Y) will be approximately constant.

(ii) A Taylor series expansion about μ is

$$h(y)=h(\mu)+(y-\mu)h'(\mu)+\frac{1}{2!}(y-\mu)^2 h''(\mu)+\dots ,$$

so

$$E\left(h(Y)\right)=h(\mu)+h'(\mu)E(Y-\mu)+\frac{1}{2}h''(\mu)E\left[(Y-\mu)^2\right]+\dots$$

$$=h(\mu)+\frac{1}{2}\sigma_Y^2 h''(\mu) \quad \text{to second order.}$$

Similarly to second order,

$$\text{Var}\left(h(Y)\right)=E\left[\left(h(Y)-E\left[h(Y)\right]\right)^2\right]=\{h'\mu\}^2 E\left[(Y-\mu)^2\right]=\sigma_Y^2\{h'(\mu)\}^2 .$$

If now $\sigma_Y=f(\mu)$, we have $\text{Var}\left(h(Y)\right)=\{f(\mu)\,h'(\mu)\}^2$ which is constant if $f(y)=\dfrac{\text{constant}}{h'(y)}$ or $\dfrac{dh(y)}{dy}\propto\dfrac{1}{f(y)}$.

(iii) Noting that all transformations include a multiplicative constant:-

If $\sigma\propto\mu$, we have $f(y)=y$ and the transformation is $\int\frac{dy}{y}=\log y$.

If $\sigma\propto\mu^2$, we have $f(y)=y^2$ and the transformation is $\int\frac{dy}{y^2}=-\frac{1}{y}$, and use $1/y$ which is the modulus.

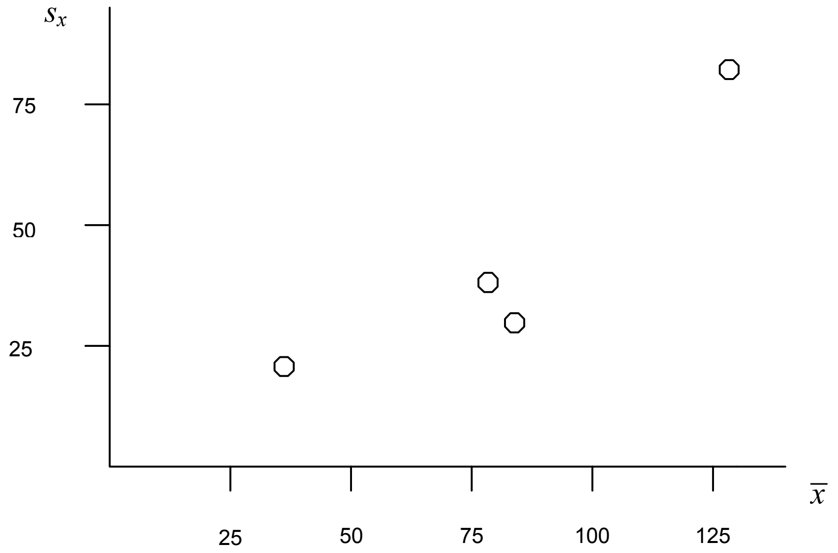
If $\sigma^2\propto\mu$, we have $f(y)=\sqrt{y}$ and the transformation is $\int\frac{dy}{\sqrt{y}}=2\sqrt{y}$, so use $\sqrt{y}$.

See next page for solution to (iv) and (v)

(iv) Descriptive statistics are

	10mg	20mg	30mg	40mg
$\bar{x}$	35.13	78.50	84.63	127.88
s_x	20.52	37.69	29.66	82.09

The standard deviation does appear to be related to the mean.



The 40mg point suggests that the form of the relation might perhaps be a curve rather than a straight line, but there is not really sufficient information to make a proper choice. A straight line would suggest the log transformation.

Calculating $\bar{x}^2/s_x, \bar{x}/s_x$ and $s_x^2/\bar{x}$ shows that $\bar{x}/s_x$ is more nearly constant than the other two ratios, suggesting that σ is approximately proportional to μ, and so log is worth trying.

(v) Possible models would be (1) a one-way analysis of variance model with amounts as treatments, each replicated 8 times, and (2) a linear regression model with x = amount. In case (1), we are not imposing a linear response of strength as amount changes. For (1), the untransformed data and the transformed data (logs) could both be analysed and the residuals studied to decide which had more nearly constant variance. For (2), residuals could be plotted against fitted values to check whether variance and expected value of response appeared to be related, again using both forms of the data. Normality of residuals could also be checked by probability plots.

If the differences between means at 10, 20, 30, 40 mg cannot be fitted satisfactorily by a linear regression model, then the one-way analysis of variance model is more satisfactory for explaining the results.

Graduate Diploma, Applied Statistics, Paper I, 2003. Question 8

(i) The "correction factor" is (grand total)$^2/N = (504)^2/54 = 4704$.

Thus the total SS is $5068 - 4704 = 364$, with 53 d.f.

$$\text{Order SS} = \frac{162^2}{18} + \frac{162^2}{18} + \frac{180^2}{18} - 4704 \;= 12, \text{ with 2 d.f.}$$

$$\text{Sex SS} = \frac{243^2}{27} + \frac{261^2}{27} - 4704 \;= 6, \text{ with 1 d.f.}$$

$$\text{Age SS} = \frac{153^2}{18} + \frac{153^2}{18} + \frac{198^2}{18} - 4704 \;= 75, \text{ with 2 d.f.}$$

The analysis of variance table can now be completed:-

Source of variation	SS	df	MS	MS ratio	
Order (O)	12.000	2	6.00	1.44	–
Sex (S)	6.000	1	6.00	1.44	–
Age (A)	75.000	2	37.50	9.00	Compare $F_{2,36}$ – significant at 0.1%
$O \times S$	61.778	2	30.89	7.41	Compare $F_{2,36}$ – significant at 1%
$O \times A$	21.667	4	5.42	1.30	Compare $F_{4,36}$ – not significant
$S \times A$	21.000	2	10.50	2.52	Compare $F_{2,36}$ – not significant
$O \times S \times A$	16.556	4	4.14	0.99	Compare $F_{4,36}$ – not significant
Residual	149.999	36	4.167		
Total	364.000	53			

(ii) The factors "Order" and "Sex" interact. The separate main effects of "Order" and "Sex" therefore have little useful meaning. There is also a main effect of "Age", and this is clearly due to the "over 30" mean being much larger than that for "under 20" and "21 – 30".

Totals and means for "Order" and "Sex" are

	TOTALS		
	1	*2*	*3*
M	68	75	100
F	94	87	80

	MEANS		
	1	*2*	*3*
M	7.56	8.33	11.11
F	10.44	9.67	8.88

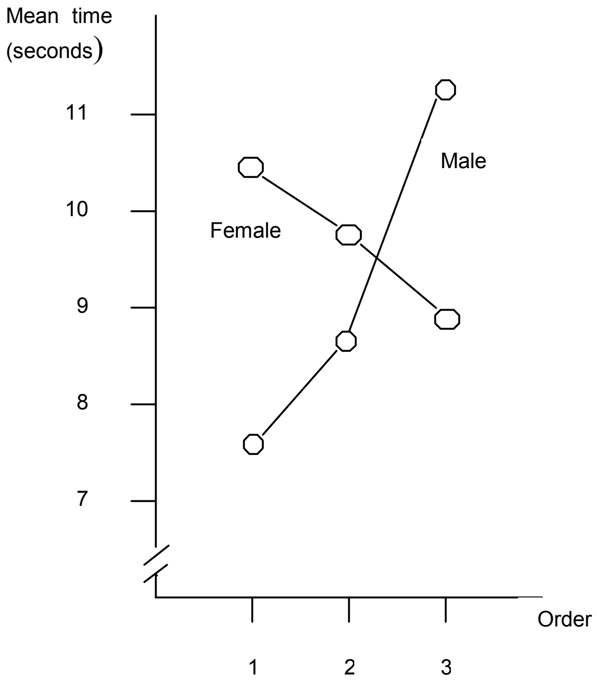
See next page for interaction diagram and continuation of solution

The "Order"–"Sex" interaction arises because males are much slower than females for order 3 whereas the opposite is true for order 1 (see diagram at foot of page). For order 2, a *t* test is appropriate:

$$\frac{9.67-8.33}{\sqrt{\frac{2\times 4.167}{9}}} = 1.39\,,$$

which is not significant as an observation from t_{36}; the difference between males and females for order 2 is not statistically significant. The same method can be used to show that the differences for orders 1 and 3 are statistically significant.

(iii) Three orders were tried for male and female subjects from three age groups. There was a clear age effect with both sexes and all orders: the average time taken by under-30s was substantially (statistically significantly) below that for over-30s. However, males and females reacted differently to the different orders. For order 1, males were slower than females; for order 2 this was true also, but the difference was not statistically significant; for order 3, females were slower than males.


Mean time
(seconds)
Male
Female
Order