THE ROYAL STATISTICAL SOCIETY

GRADUATE DIPLOMA EXAMINATION

NEW MODULAR SCHEME

introduced from the examinations in 2009

MODULE 4

SOLUTIONS FOR SPECIMEN PAPER B

THE QUESTIONS ARE CONTAINED IN A SEPARATE FILE

The time for the examination is 3 hours. The paper contains eight questions, of which candidates are to attempt **five**. Each question carries 20 marks. An indicative mark scheme is shown within the questions, by giving an outline of the marks available for each part-question. The pass mark for the paper as a whole is 50%.

The solutions should not be seen as "model answers". Rather, they have been written out in considerable detail and are intended as learning aids. For this reason, they do not carry mark schemes. Please note that in many cases there are valid alternative methods and that, in 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 the questions and solutions, the Society will not be responsible for any errors or omissions.

The Society will not enter into any correspondence in respect of the questions or solutions.

Note. In accordance with the convention used in all the Society's examination papers, the notation log denotes logarithm to base *e*. Logarithms to any other base are explicitly identified, e.g. $\log_{10}$.

Graduate Diploma Module 4, Specimen Paper B. Question 1

(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

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

so

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

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

Similarly to second order,

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

If now $\sigma_Y = f(\mu)$, we have $\mathrm{Var}\left(h(Y)\right) = \left\{f(\mu)h'(\mu)\right\}^2$ 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.

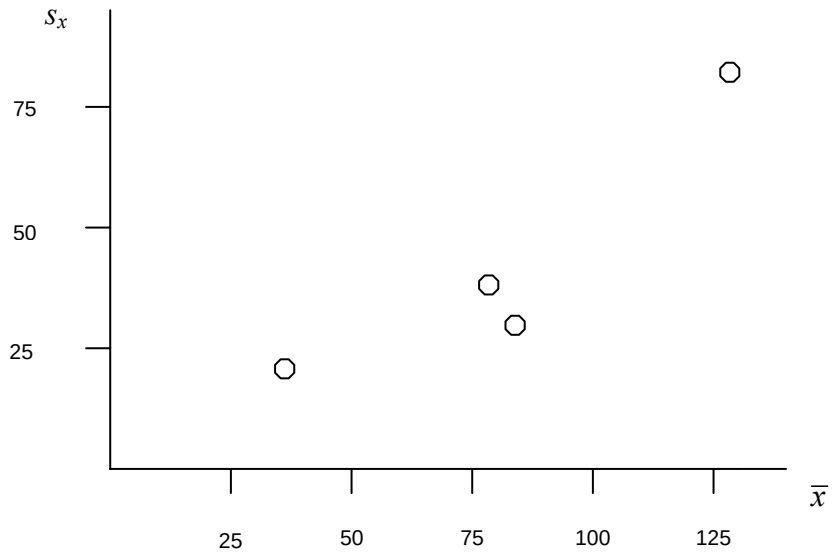
If $\sigma^2 \propto \mu$, we have $f(y) = \sqrt{y}$ and the transformation is $\int \frac{dy}{\sqrt{y}} = 2\sqrt{y}$, so use $\sqrt{y}$.

Solution continued on next page

(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 Module 4, Specimen Paper B. Question 2

(i) (a) The probability function of the binomial distribution B(m, π) written in exponential form is

$$f(y,\pi) = \exp\left[\log\binom{m}{y} + y\log\pi + m\log(1-\pi) - y\log(1-\pi) \right]$$

so $\log f(y_i,\pi_i) = y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + m_i \log(1-\pi_i) + \text{constant}$.

As y_i is on the right of this equation, it is in canonical form, and the multiplier of y_i is the natural parameter, which is therefore

$$\log\left(\frac{\pi_i}{1-\pi_i}\right).$$

[This is the log-odds, or logit.] The generalised linear model sets this link function equal to the linear predictor.

(b) The *odds* is the ratio of probabilities of "success" and "failure" for Y_i, i.e. $\pi_i/(1-\pi_i)$. The *log-odds* is simply the logarithm (base e) of this, as used in the link function.

After the generalised linear model has been fitted, the (estimated) value of η_i is obtained – this is the estimate of the log-odds. We have

$$\eta_i = \log\left(\frac{\pi_i}{1-\pi_i}\right) \quad \Rightarrow \quad \frac{\pi_i}{1-\pi_i} = \exp(\eta_i)$$

so the estimate of the odds is $\exp(\eta_i)$.

Given the necessary standard errors (SE), approximate 95% confidence limits are "estimate ± 1.96×SE" for each of odds and log-odds. The details are in (ii)(c) below.

(ii) (a) We are not told how the sampling was carried out, so the independence of observations is not guaranteed; neither is the randomness.

The analysis would be appropriate if a random sample of data from a larger population has been selected, omitting multiple births (twins etc) and only using a mother once if she has had more than one child at different times [to avoid probable lack of independence]. Many hospitals would need to be represented in the sampling, as well as home births. It would not be appropriate to use this analysis if the "group of women" mentioned came from a limited area, for example by studying all births from the local hospital over a few years.

Solution continued on next page

(b) Step 1 chooses the single predictor variable which reduces the scaled deviance as much as possible from the "constant only" model. Clearly this is GEST, the length of gestation period, which reduces the deviance by 339.37 (on 1 df). We next consider adding AGE, and this step further reduces the deviance by 6.566, also on 1 df; this is significant as an observation from χ^2 with 1 df, so AGE should be included. So we use AGE and GEST in the model.

(c) The coding AGE = 0, GEST = 0 gives

$$\hat{\eta} = -1.7659, \quad \mathrm{SE}(\hat{\eta}) = 0.1296 .$$

The estimate of the odds is exp(–1.7659) = 0.171.

95% confidence limits for η are –1.7659 ± 1.96×0.1296, i.e. (–2.020, –1.512), so the corresponding limits for the odds are (0.1327, 0.2205) after exponentiating.

The estimate of the probability of mortality is

$$\hat{\pi} = \frac{e^{\hat{\eta}}}{1+e^{\hat{\eta}}} = \frac{0.171}{1.171} = 0.146 .$$

Similarly, the upper and lower limits of the 95% confidence interval for this probability are as follows.

$$\text{Lower limit: } \frac{0.1327}{1.1327} = 0.117 ; \quad \text{upper limit: } \frac{0.2205}{1.2205} = 0.181 .$$

(d) GEST is now to be coded as 1, AGE remaining zero. The log-odds ratio for this group compared to the group in (c) is thus simply the value of the GEST parameter, i.e. –3.2886.

So the odds ratio is exp(–3.2886) = 0.0373.

95% confidence limits for this log-odds ratio are –3.2886 ± 1.96×0.1846, i.e. (–3.650, –2.927). Thus the limits for the odds ratio are exp(–3.650) = 0.026 and exp(–2.927) = 0.054.

Graduate Diploma Module 4, Specimen Paper B. Question 3

(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.

Solution continued on next page

(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 Module 4, Specimen Paper B. Question 4

(a) (i) A factor is a categorical variable in which values are simply codes for each category, as in types of house. A continuous variable is an observation recorded on a scale on which any real value (within some range) is possible.

(ii) There are 2 d.f. for regression, and both "age" and "type" are used as regressor variables. Thus "type" must have been treated as a continuous variable because 2 d.f. would be needed for a factor variable with 3 levels, leaving none for age. Type of house could be regarded as a proxy for "number of detached sides", but there seems no good reason to assume a linear scale for it. Factor coding would be better.

(iii) Package B: $\begin{matrix} 1 & 58 & 0 & 0 \\ 1 & 19 & 0 & 1 \\ 1 & 10 & 1 & 0 \end{matrix}$ would be the first three rows of the design matrix.

The coefficient of "age" is the same in each package, –0.4180.

For type 1, A gives 78.8603 + 11.2249 + 0 = 90.0852
and B gives 68.212 + 21.873 + 0 = 90.085 (same)

For type 2, A gives 78.8603 – 0.5764 = 78.2839
and B gives 68.212 + 10.072 = 78.284 (same)

For type 3, A gives 78.8603 – 11.2249 + 0.5764 = 68.2118
and B gives 68.212 = 68.212 (same)

ANOVA tables will show identical values for DF, SS, MS, F and p. SEs, p-values and confidence intervals for coefficients for factor and constant will be different; those for the continuous variable "age" will be identical. (Type is in fact roughly linear as the results show.)

(b) With 62 d.f. all the given correlation coefficients are significant (at 1%). Scales A, B are strongly negatively correlated – fear of falling goes with lack of confidence doing risky tasks. The anxiety scale C is positively related to A, as would be expected, and is rather weakly opposed to B – very anxious people have less confidence in undertaking tasks.

Thus if a simple linear regression of B on C were to be calculated, the regression coefficient would be negative. But in the multiple regression of B on A and C, there is already a component of anxiety modelled by scale A, and partial correlations are required to give the complete picture of relationships.

Graduate Diploma Module 4, Specimen Paper B. Question 5

(i) An appropriate model is

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk},$$

In this model:

μ is the grand mean

α_i refers to machine i (i = 1, 2, 3). This is a fixed effect since we are interested only in these three machines; thus $\sum_{i=1}^{3} \alpha_i = 0$

β_j refer to employee j (j = 1, 2, …, 6). This is a random effect since the employees were chosen at random from all those available. The $\{\beta_j\}$ are uncorrelated with mean 0 and variance σ_B^2

$\{(\alpha\beta)_{ij}\}$ are interactions. For each level of A (i = 1, 2, 3), they are uncorrelated with one another, with $\{\varepsilon_{ijk}\}$ or with $\{\beta_j\}$, and have mean 0 and variance σ_{AB}^2. For each level of B (j = 1 to 6) the $\{(\alpha\beta)_{ij}\}$ are constants and $\sum_i (\alpha\beta)_{ij} = 0$

$\{\varepsilon_{ijk}\}$ are random variables, uncorrelated with one another, with $\{\beta_j\}$ or $\{(\alpha\beta)_{ij}\}$ and with mean 0 and variance σ^2

(ii) $E[MS_A] = \sigma^2 + 2\sigma_{AB}^2 + 6\sum_{i=1}^{3} \alpha_i^2$

$E[MS_B] = \sigma^2 + 6\sigma_B^2$

$E[MS_{AB}] = \sigma^2 + 2\sigma_{AB}^2$

Solution continued on next page

(iii) The (corrected) total SS is given in the question: 2071.99.

$$\text{SS machines} = \frac{656.8^2}{12} + \frac{710.7^2}{12} + \frac{797.1^2}{12} - \frac{2164.6^2}{36}$$

$$= 130987.4283 - 130152.5878 = 834.84$$

$$\text{SS employees} = \frac{365.8^2}{6} + \frac{346.4^2}{6} + \frac{383.5^2}{6} + \frac{359.8^2}{6} + \frac{401.7^2}{6} + \frac{307.4^2}{6} - \frac{2164.6^2}{36}$$

$$= 131031.4233 - 130152.5878 = 878.84$$

Thus, using the information in the question,

$$\text{SS interaction} = \frac{264308.12}{2} - 130152.5878 - 834.84 - 878.84 = 287.79$$

and

residual SS = 2071.99 – (sum of all above SSs) = 70.52.

Hence the analysis of variance is

Source of variation	d.f.	Sum of squares	Mean square	*F* value
Machines (M)	2	834.84	417.42	417.42/28.78 = 14.50
Employees (E)	5	878.84	175.77	175.77/3.92 = 44.86
M × E interaction	10	287.79	28.78	28.78/3.92 = 7.35
Residual	18	70.52	3.92	
Total	35	2071.99		

To test the null hypothesis "all $\alpha_i = 0$", 14.50 is referred to the $F_{2,10}$ distribution. This is significant at the 1% level, so there is strong evidence against this null hypothesis. We may assume that there are differences among the means for machines; machine 3 is the best to buy.

There is however very strong evidence of an interaction between machines and employees; 7.35 is significant at the 0.1% level when referred to $F_{10,18}$, so we reject the null hypothesis that $\sigma^2_{AB} = 0$.

However, the table of totals shows that machine 3 is best for <u>all</u> employees, though less so for some employees than for others. So machine 3 still appears best overall. It appears that machine 2 is better than machine 1 for some employees but not for others.

There is also very strong evidence (refer 44.86 to $F_{5,18}$) that there are differences within the population of employees (the null hypothesis that $\sigma_B{}^2 = 0$ is rejected).

Graduate Diploma Module 4, Specimen Paper B. Question 6

An "incomplete block" scheme of some sort is needed when block size (the number of "plots" in a "block") is less than the number of treatments to be compared. It is obviously not possible in such circumstances for every treatment to appear in every block. A balanced incomplete block is a design where a degree of symmetry is nevertheless preserved, and is useful when it is desired that comparisons between each pair of treatments are to be made with the same precision. It requires all blocks to be the same size. The design is such that every pair of treatments occurs together in the blocks the same number of times. This is illustrated in the example in this question, where there are 7 treatments in blocks of size 3, and each pair of treatments occurs together in the blocks just once.

If the blocks cannot all be the same size, or if some comparisons are more important than others, less balanced designs may be necessary or preferred.

(i) This is a balanced incomplete block design (see discussion above) with structural parameters as follows.

N is the number of observations: $N = 21$.

b is the number of blocks: $b = 7$.

k is the size of each block: $k = 3$.

v is the number of treatments: $v = 7$.

r is the number of replicates of each treatment: $r = 3$.

λ is the number of times each pair of treatments occurs together in a block: $\lambda = 1$. [Note: $\lambda = r(k-1)/(v-1) = 3\times2/6 = 1$; this has to be an integer for the incomplete block design to be *balanced*.]

(ii) The total SS is $8341 - \dfrac{413^2}{21} = 218.667$.

The SS for blocks is

$$\frac{1}{3}\left(\frac{59^2}{3}+\frac{66^2}{3}+\ldots+\frac{63^2}{3}\right)-\frac{413^2}{21} = 90.000.$$

The SS for treatments adjusted for blocks is [formula quoted in question]

$$\frac{7\times1}{3}\sum_i \hat{\tau}_i^{\,2} = \frac{7}{3}\times 41.3908 = 96.579.$$

Solution continued on next page

Hence:

SOURCE	DF	SS	MS	F value
Blocks	6	90.000	–	
Treatments adjusted for blocks	6	96.579	16.097	4.01
Residual	8	32.088	4.011	$= \hat{\sigma}^2$
TOTAL	20	218.667		

The F value of 4.01 is referred to $F_{6,8}$; this is significant at the 5% level (critical point is 3.58), so there is some evidence that there are differences between the treatments, having adjusted for the blocks. The differences are explored in part (iii).

(iii) The variance of the difference between any pair of treatment means is estimated by [formula quoted in question]

$$\frac{2k\hat{\sigma}^2}{v\lambda} = \frac{2\times 3\times 4.011}{7\times 1} = 3.438\,.$$

Least significant differences are therefore given by $t\times\sqrt{3.438}$ where t denotes the appropriate critical point from the t_8 distribution: 2.306 for 5%, 3.355 for 1%, 5.041 for 0.1%. So the respective LSDs are 4.28, 6.22, 9.35.

The table below shows the estimated treatment effects (adjusted for blocks) in ascending order of size.

E	D and F	A	C	B	G
–2.8571	–1.8571	–0.2857	–0.1429	2.5714	4.4290

Interpretation is difficult. Recall that the overall test in the analysis of variance in part (ii) was only significant at the 5% level. In LSD terms, we see that *all* the treatments could be considered the same if judged at the 0.1% level. At the 1% level, G is "detached" from (and better than) E and (D and F), but no better than A, C or B which are themselves no better than E or (D and F). At the 5% level, G is "detached" from (better than) all but B, while B is also "detached" from E and (D and F).

Graduate Diploma Module 4, Specimen Paper B. Question 7

A contrast among treatment means is $\sum c_i \bar{y}_i$, where the c_i are a set of constants whose sum is zero. Usually the c_i are integers, for simplicity in calculations.

If the variance of individual observations is σ^2, then that of the mean $\bar{y}_i$ is σ^2/r. The variance of the contrast $\sum c_i \bar{y}_i$, assuming independence of all observations (proper randomisation) is $\sum c_i^2 \mathrm{Var}(\bar{y}_i)$, which is $\sum c_i^2 \sigma^2 / r$. The standard deviation is the square root of this, and thus the standard error is $\sqrt{\sum c_i^2 s^2 / r}$ where s^2 denotes the residual mean square which estimates σ^2.

Two orthogonal contrasts among the same set of means have coefficients c_i and d_i such that $\Sigma c_i = 0 = \Sigma d_i$ and $\Sigma c_i d_i = 0$. The importance of orthogonal contrasts is that the are uncorrelated. Thus they are independent for the case of Normally distributed errors, and represent comparisons among the means that can be independently estimated and tested for.

(i) and (ii)

The required contrasts are

	A	B	C	D	E	F	G	H
Mean	88	198	66	235	265	233	40	41
(a)	–1	1	–1	1	–1	1	–1	1
(b)	–1	–1	1	1	0	0	0	0
(c)	1	1	1	1	1	1	–3	–3
(d)	1	1	1	1	–2	–2	0	0
(e): (b) with (a)	1	–1	–1	1	0	0	0	0
(e): (c) with (a)	–1	1	–1	1	–1	1	3	–3
(e): (d) with (a)	–1	1	–1	1	2	–2	0	0

We have $s^2/r = 3265.8/5 = 653.16$. The SE for each contrast is thus $\sqrt{653.16 \Sigma c_i^2}$. The number of df for the residual is 39 – 7 = 32 (there are 40 observations and 8 treatments). So the statistical significance of each contrast may be tested by referring (value)/SE to the t distribution with 32 df, on the assumption of Normality and common variance for the experimental errors. The two-sided critical points of t_{32} are 2.037 for 5%, 2.738 for 1% and 3.622 for 0.1%.

Solution continued on next page

We have, from the above table,

	Value	$\sum c_i^2$	SE	Value/SE
(a)	248	8	72.29	3.431
(b)	15	4	51.11	0.293
(c)	842	24	125.20	6.725
(d)	–409	12	88.53	–4.620
(b) with (a)	59	4	51.11	1.154
(c) with (a)	244	24	125.20	1.949
(d) with (a)	343	12	88.53	3.874

There is strong evidence of an overall difference between the effects of the levels of the fertiliser [contrast (a)]; it appears that high fertiliser level is better than low level.

There is no evidence of difference between the effects of the cultures [contrast (b)].

There is very strong evidence for an effect of inoculation [contrast (c)]; it appears that inoculation gives higher yield.

Likewise there is very strong evidence for an effect of the two strains of Rhizobium [contrast (d)]; it appears that CC 511 performs better than R 3644.

However, interpretations must take account of any interactions. There is no evidence of interaction between the two cultures of R3644 and fertiliser level [(b) with (a)]. There is also not (quite) sufficient evidence to suggest an interaction between the effect of inoculation and fertiliser level [(c) with (a)]. There is very strong evidence of an interaction between the two strains of Rhizobium and fertiliser level [(d) with (a)]: it appears that R 3644 performs better at the high fertiliser level than at the low level, but CC 512 somewhat better at the low fertiliser level than the high.

Graduate Diploma Module 4, Specimen Paper B. Question 8

Part (i)

← N	I				II	III	IV	S →
DOOR	*ab*	(1)	*bc*	*b*				DOOR
	c	*abc*	*a*	*ac*				

There is likely to be a "climatic trend" from north to south, even in a glasshouse, increased by having doors at each end which will produce temperature changes when opened. Blocking in this direction, as shown, is therefore a good property of the design. The eight treatment combinations will be randomised in each block, independently of one another.

Part (ii)

(a) The remaining sums of squares are calculated as follows. We need the grand total, 304.0, and hence the "correction factor" $304.0^2/32 = 2888$.

$$\text{SS for blocks} = \frac{68.8^2}{8} + \frac{81.8^2}{8} + \frac{83.3^2}{8} + \frac{70.1^2}{8} - 2888 = 2909.6975 - 2888 = 21.6975 .$$

$$\text{SS for } A = \frac{(217.1 - 86.9)^2}{32} = 529.75125 .$$

$$\text{SS for } ABC = \frac{(155.8 - 148.2)^2}{32} = 1.80500 .$$

(We may check that the sums of squares for all seven main effects and interactions add up to the stated treatments total of 616.795.)

By subtraction, the residual SS = total SS – treatments SS – blocks SS = 36.7875.

Each main effect and interaction has 1 degree of freedom, giving 7 in all for the treatments, and the residual has 21.

Solution continued on next page

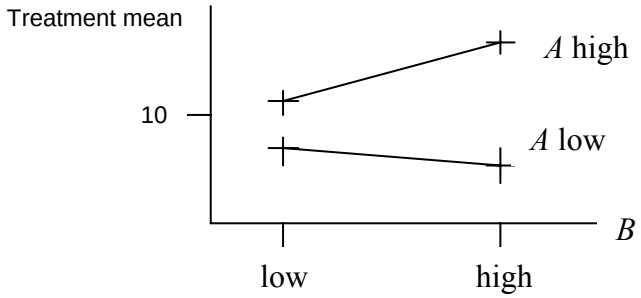
Hence:

SOURCE	DF	SS	MS	*F* value
Blocks	3	21.6975	7.2325	4.13 compare $F_{3,21}$
A	1	529.75125	529.75125	302.40 compare $F_{1,21}$
B	1	19.84500	19.84500	11.33 …
C	1	1.05125	1.05125	0.60 …
AB	1	62.16125	62.16125	35.48 …
AC	1	0.98000	0.98000	0.56 …
BC	1	1.20125	1.20125	0.69 …
ABC	1	1.80500	1.80500	1.03 …
Treatments	7	616.7950		
Residual	21	36.7875	1.7518	$= \hat{\sigma}^2$
TOTAL	31	675.2800		

(b) $F_{1,21}$ tests for the main effects and interactions (upper 5% point is 4.32) show that *A*, *B* and *AB* are significant. The blocks effect is also significant (upper 5% point of $F_{3,21}$ is 3.07).

To study the effects of *A* and *B* in the presence of an *AB* interaction, we need the table of *AB* means:

	A low	*A* high	[Treatments included]	
B low	6.04	11.39	(1), *c*	*a*, *ac*
B high	4.83	15.75	*b*, *bc*	*ab*, *abc*



(c) The decision to include blocking was wise. There is evidence that the blocks are not all the same as each other, and we see that the two end blocks performed less well than those in the centre.

Because the main effect of *C* was not significant and there were no interactions involving *C*, we may conclude that it does not matter which of the two experimental levels of *C* is used in practice. We may, of course, still wish to explore higher or lower levels in a later experiment.

Factors *A* and *B* interact, so their main effects should not be examined alone. Instead, we refer to the table of *AB* means. Clearly use of the high level of *A* has a beneficial effect, and this is increased by using the high level of *B*. On the other hand, *A* does not perform well at the low level and is even worse at this level if the high level of *B* is used.