ROYAL
STATISTICAL
SOCIETY
DATA | EVIDENCE | DECISIONS

THE ROYAL STATISTICAL SOCIETY

2016 EXAMINATIONS − SOLUTIONS

GRADUATE DIPLOMA – MODULE 4

The Society is providing these solutions to assist candidates preparing for the examinations in 2017.

The solutions are intended as learning aids and should not be seen as "model answers".

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.

Question 1

(i) The least squares estimator is given by $\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ (1 mark). The Gauss-Markov theorem states that these estimators have minimum variance (1 mark) among all such linear unbiased estimators (1 mark).

(ii) Multicollinearity refers to a situation where two or more explanatory variables are (nearly) perfectly linearly related (1 mark), i.e. where two explanatory variables are highly correlated (0.5 marks), or where one explanatory variable is a (nearly) perfect linear combination of two or more other explanatory variables (0.5 marks). The first scenario can be detected by considering the correlations between pairs of explanatory variables, where a correlation near to -1 or+1 would indicate near multicollinearity (1 mark). However, as the second case is the relationship among 3 or more variables, this cannot be detected by considering pairwise correlation coefficients, but requires assessment through multiple linear regression of one explanatory variable on the remaining explanatory variables (1 mark).

(iii) Stepwise regression methods provide a way of searching through possible explanatory models without having to consider all possible models (1 mark). Forwards selection methods start with the null model and at each step identify the variable that explains the largest proportion of the unexplained variability. If a statistic for including this "best" variable exceeds whatever threshold is set for inclusion of a variable, then the term is added to the current model, and the process is repeated until there are no further variables that could added to the model to improve it (1 mark). Backwards elimination methods start with the full model and at each step identify the variable that explains the smallest proportion of the explained variability. If the statistic for excluding the "worst" variable fails to exceed whatever threshold is set for exclusion of a variable, then the variable is removed from the current model, and the process is repeated until there are no further variables that could be removed from the model without making it worse (1 mark). Choice of variance ratio thresholds will impact on how likely it is for variables to be added to or removed from the current model at each step (1 mark).

(iv) In the first step, DiamB is removed from the full model as this has the last impact on the quality of the fit, with a MS Ratio considerable smaller than the threshold of 1.00 (1 mark). In the second step, AngB is removed from the model as this has little impact on the quality of the fit, with a MS Ratio also considerable smaller than the threshold of 1.00 (1 mark). Both of these removals can be anticipated based on the size of the correlation coefficients with CW. In the third step Ht is removed from the model, with a MS ratio just smaller than the threshold of 1.00, despite having a fairly strong correlation with CW. This is because of the partial correlation of Ht with DBH, which has a stronger correlation with CW – so some of the variation in CW that could be explained by HT is also explained by DBH (1 mark). The chosen model indicates that CW increases significantly with increasing values of DBH (by 0.22 units for every increase of 1 unit) and increases at a smaller rate with decreasing values of HLC (0.03 units for every decrease of 1 unit) (as would be anticipated from the correlation matrix) (1 mark). Hence the crown width (maximum width of foliage) is larger when the diameter of the trunk is larger and smaller as the height of the crown increases – tall thin trunks have smaller canopies than short fat trunks (1 mark).

(v) The assumptions behind the analysis are that the residuals are independent (0.5 marks), Normally distributed with mean zero (0. 5 marks), and with a constant variance (0. 5 marks), and that the relationships between the response variable and each of the explanatory variables are linear (0. 5 marks). The residuals are calculated as the difference between the observed values and the values fitted for the chosen model (0. 5 marks). Plotting the residuals against the fitted values provides a graphical method of assessing the homogeneity of variance assumption (looking for equal scatter of the residuals about zero across the range of fitted values) (0. 5 marks), a histogram of the residuals or a plot of the ordered residuals against the Normal quantiles provides an assessment of the Normality assumption (0. 5 marks), and plotting the residuals against each of the explanatory variables provides an assessment of the linearity assumption (0. 5 marks).

Question 2

(i) The systematic variation that is to be eliminated may arise from differences between the rows (running across the field, different moisture levels) and columns (running from top to bottom, different previous nitrogen fertilizer treatments) of the arrangement of the plots (1 mark). In the Latin square design these effects are assigned to the rows and columns of the square, with the treatments appearing in the square such that each treatment appears exactly once in each row and once in each column (1 mark). Thus, if there are any systematic differences between the rows, they will affect each of the treatments equally, and likewise for the columns. The rows, columns and treatments are orthogonal in this sense. It has to be assumed that there is no interaction between rows, columns and treatments (1 mark).

(ii) The “correction term” is $\frac{118.98^2}{25}$ = 566.25 (0.5 marks), so the total sum of squares is 571.69 – 566.25 = 5.44 (0.5 marks).

The Row SS is $\frac{21.39^2}{5}+\frac{22.02^2}{5}+\frac{24.35^2}{5}+\frac{24.68^2}{5}+\frac{26.54^2}{5}-566.25 = 569.76-566.25 = 3.51$ (0.5 marks).

The Column SS is $\frac{24.58^2}{5}+\frac{23.70^2}{5}+\frac{24.37^2}{5}+\frac{23.47^2}{5}+\frac{22.86^2}{5}-566.25 = 566.64 - 566.25 = 0.39$ (0.5 marks).

The Sulphur SS is $\frac{22.92^2}{5}+\frac{24.12^2}{5}+\frac{24.98^2}{5}+\frac{24.20^2}{5}+\frac{22.76^2}{5}-566.25 = 566.95 - 566.25 = 0.70$ (0.5 marks).

Hence (0.5 marks for the degrees of freedom, 1 mark for the mean squares, 1 mark for the F-values):

SOURCE	DF	SS	MS	F-value
Rows	4	3.51	0.878	12.54
Columns	4	0.39	0.098	1.40
Sulphur	4	0.70	0.175	2.50
Residual	12	0.84	0.070	
TOTAL	24	5.44		

The critical values for $F_{4,12}$ are 2.48 (p = 0.10), 3.26 (p = 0.05), 5.41 (p = 0.01) and 9.63 (p = 0.001) (0.5 marks). Interpretation is that the Row effect is very highly significant (P<0.001), indicating the importance of blocking for differences in moisture levels from the top to bottom of the field (0.5 marks), but that the Column effect is not significant (P>0.1), indicating that the previous nitrogen fertilizer treatments did not appear to influence this new response (0.5 marks). The Sulphur effect is only just significant at the 10% level, so there is little evidence for an overall effect, but the quantitative nature of this factor might justify some further exploration (0.5 marks). *The full 2 marks for the interpretation should be awarded for a clear description of the importance of the effects, supported by appropriate critical values and/or identification of the appropriate significance levels. 1 mark should be lost for failing to state either critical values or significance levels.*

(iii) The mean response at 0 kg is 4.584 and that at 10kg is 4.824, so that the difference in means is 0.240 (0.5 marks). Based on the residual mean square from the ANOVA, the SED is $\sqrt{\frac{2*0.070}{5}} = 0.167$ (0.5 marks) so that a 95% confidence interval is 0.240 +/- (2.179*0.167) = (-0.124, 0.604) (1 mark).

(iv) Orthogonal contrasts lead to independent terms in an analysis of variance, and hence to independent tests and comparisons (1 mark). Further, the sum of squares for a complete set of orthogonal contrasts add up to the total treatment sum of squares in the analysis of variances, and this, by testing each one against the residual mean square in the usual way, gives a mechanism for investigating where any treatment differences lie, including investigating the shape of the response to a quantitative input (1 mark). The orthogonality of a pair of contrasts can be assessed by calculating the sum of the products of the coefficients for the contrasts – if the sum is zero, then the contrasts are orthogonal (1 mark).

(v) The linear contrast value is $\frac{(-2)*22.92+(-1)*24.12+(0)*24.98+(+1)*24.20+(+2)*22.76}{5} = \frac{-0.24}{5} = -0.048$ (1 mark – the divisor of 5 allows for the replication level, but the contrast value can be marked as correct as -0.24 without this adjustment). So the sum of squares for this contrast is $\frac{(-0.24)^2}{((-2)^2+(-1)^2+(0)^2+(+1)^2+(+2)^2)*5} = 0.001$ (0.5 marks). This gives an F-value of 0.01 which is not significant (P>0.1), so there is no evidence for an overall linear increase in grain yield in response to increasing sulphur (0.5 marks).

The quadratic contrast value is $\frac{(+2)*22.92+(-1)*24.12+(-2)*24.98+(-1)*24.20+(+2)*22.76}{5} = \frac{-6.92}{5} = -1.384$ (1 mark – the divisor of 5 adjusts for the replication level, but the contrast value can be marked as correct as -6.92 without this adjustment). So the sum of squares for this contrast is $\frac{(-6.92)^2}{((+2)^2+(-1)^2+(-2)^2+(-1)^2+(+2)^2)*5} = 0.684$ (0.5 marks). This gives an F-value of 9.77 which is significant at the P<0.01 level, with grain yield increasing at low levels of sulphur input but then declining (above the 20kg level) (0.5 marks).

The remaining sum of squares for the Sulphur effect after allowing for these two contrasts is less than 0.02 on 2 degrees of freedom, giving a mean square of less than 0.01 and an F-value of around 0.15, which is not significant (P>0.1) indicating that there is no evidence for lack of fit around the quadratic model. (1 mark)

Question 3

(i) The advantages of using a randomised complete block design are that the analysis and interpretation are simpler, with each treatment occurring in each block, and all pairwise comparisons of treatments occurring in each block (1 mark). The disadvantage is that each block would need to use material produced from more than one run of the initial step, potentially increasing the within-block variability (1 mark). The advantage of using an incomplete block design is that each block could then be associated with material from a single run of the initial step, so that the blocking structure matches the natural structure of the experimental units (1 mark). The disadvantage of using an incomplete block design is that it might not be possible for all pairwise comparisons of treatments to occur equally often in blocks (a balanced design), so that differences between blocks might influence observed differences between treatments (1 mark).

(ii) The first relationship (rt = bk) relates to the size of the experiment, so that the number of treatments multiplied by the number of replicates must be equal to the number of blocks multiplied by the number of units per block (1 mark). The second relationship relates to the number of pairwise comparisons for each treatment, λ(t-1), which must be equal to the number of possible within-block comparisons for each treatment, r(k-1) (1 mark).

(iii) For t=9 and r=8, we first use the second relationship to identify the block size, k, for each value of λ. With this value of k we then use the first relationship to calculate the number of blocks, b, where all parameters must have integer values. (1 mark)

(a) For λ=1, we have 1*(9-1) = 8*(k-1), so that k=2 (0.5 marks). We then have 8*9 = 2*b, so that b=36 (0.5 marks)

(b) For λ=2, we have 2*(9-1) = 8*(k-1), so that k=3 (0.5 marks). We then have 8*9 = 3*b, so that b=24 (0.5 marks)

(c) For λ=5, we have 5*(9-1) = 8*(k-1), so that k=6 (0.5 marks). We then have 8*9 = 6*b, so that b=12 (0.5 marks)

For λ=4, we have 4*(9-1) = 8*(k-1), so that k must be equal to 5. We then have 8*9 = 5*b, but the solution for b is not then an integer (b = 14.4), so that a balanced incomplete block design does not exist for the specified combination of parameters. (1 mark)

(iv) As indicated in the hints provided, the design is constructed in two halves, each containing 4 replicates of the 9 treatments, and each constructed using a cyclic approach. In each set, twelve within-block comparisons are available for each treatment, most equally divided between the eight possible comparisons so that each treatment appears once in a block with four of the other treatments, and twice in a block with the remaining four treatments (1 mark). The cyclic construction approach starts by allocating the first replicate of the treatments in standard order to the nine blocks and then shifting the set of treatments forward or back different numbers of blocks, carefully choosing the sizes of the shifts to obtain the required patterns of within block joint occurrences (1 mark). For one set of nine blocks, shifting the second replicate back by 1 block and the third replicate back by 3 blocks provides within-block comparisons of each treatment with 6 of the other 8 treatments (1 mark – other combinations that achieve this same pattern are also possible and acceptable solutions). The shift for the fourth replicate then needs to ensure that each treatment occurs within blocks with all 8 of the other treatments, and can be determined by considering the within-block comparisons achieved for a particular treatment in these first three replicates (0.5 marks). As treatment 1 occurs within blocks with treatments 2, 4, 7, 8, 9 and 3 in the above solution, it now needs to appear with treatments 5 and 6, and so should be added to block 5 in which both of these treatments occur (so shifting the fourth replicate back by 5 blocks (0.5 marks).

1	2	3	4	5	6	7	8	9
2	3	4	5	6	7	8	9	1
4	5	6	7	8	9	1	2	3
6	7	8	9	1	2	3	4	5

Treatment 1 occurs twice in blocks with treatments 3, 5, 6 and 8 and once with the other four, treatment 2 with treatments 4, 6, 7 and 9, and so on.

For the other set of nine blocks, a similar approach is taken, first shifting the second and third replicates forward by 1 and 3 blocks respectively, giving a different set of within-block comparisons to those for the first set (0.5 marks – different shifts will be appropriate here if different choices were made for the first set of nine blocks). The shift for the fourth replicate is then chosen so that each treatment occurs within blocks with all eight other treatments, and twice with the half of the treatments with which it only occurred once in the first set of blocks (0.5 marks). As treatment 1 again still needs to occur with treatments 5 and 6, this can be achieved by shifting either 4 or 5 blocks forward (0.5 marks – a different set of options would occur if different choices are made above) – shifting 4 blocks forward clearly gives second within-block occurrences with treatments 2 and 4, while shifting 5 blocks forward gives second within-block occurrences for treatment 3, and so the 4-block shift provides the complementary sets of comparisons to the first set of nine blocks (0.5 marks – a different decision would be appropriate if different choices are made above).

1	2	3	4	5	6	7	8	9
9	1	2	3	4	5	6	7	8
7	8	9	1	2	3	4	5	6
6	7	8	9	1	2	3	4	5

If a cyclic construction approach has been used, the checks described above for the construction should ensure that the constructed design has the appropriate properties of equal occurrence of all pairs of treatments within blocks. The same design could be *constructed without taking a cyclic approach, but this would require the demonstration of the pattern of within-block pairwise occurrences within each set of nine blocks. Full marks should be awarded if a complete design is constructed using an alternative approach, with a clear description of the approach – should involve the checking of joint occurrences as treatments are added to blocks, to meet the criterion described for the first mark of this part of the question. Where a non-cyclic approach has been taken, a maximum of 3 marks should be awarded for a partial solution where the construction has been approached by sequentially adding treatments to blocks to satisfy this criterion, and is clearly described. A solution without dividing the blocks into two sets of nine is very difficult to achieve, but again should be awarded a maximum of 3 marks where a clear description of the approach is provided – based on the sequential addition of treatment to blocks to satisfy the criterion described.*

(v) One set of nine blocks should be processed in the first three days, and the other set in the last three days, so a first randomisation step is to allocate sets of blocks to sets of three days (0.5 marks). The second randomisation step is to allocate three blocks in a randomised order to each day within these sets (0.5 marks). The third randomisation step is to randomise the order in which the treatment codes are processed within each block (1 mark), and the fourth randomisation step is to randomise the allocation of the nine treatments, A to I, to the treatment codes (1 mark). *An alternative randomisation approach, where the eighteen blocks are randomly allocated in groups of 3 to the 6 days, without considering the two sets of 9 blocks in the construction, is acceptable though less obvious (and could result in a bigger between day effect due to the same treatments occurring together in blocks), but should still receive the 1 mark allocated to the first two steps described above.*

Question 4

(i) Linear models are such that the response is a linear function of the parameters (so includes curvilinear functions such as quadratic or cubic polynomial responses) (1 mark). Non-linear models do not allow the response to be expressed as a linear combination of the parameters (1 mark), though will usually include some parameters to which the response is linearly related (1 mark). Examples include exponential and logistic functions (0.5 marks). The Newton-Raphson procedure uses an iterative numerical approach to find the optimal parameter values for a non-linear regression model (1 mark), as an algebraic solution is not available (0.5 marks). The approach uses the curvature of the "surface" relating the residual sum of squares to the parameter values to find successively better approximations to the optimal solution (1 mark).

(ii) The models are:

(a) Linear in all four parameters (*a*, *b*, *c*, *d*) (1 mark)

(b) Non-linear with respect to parameters *b* and *m*, but linear with respect to parameters *a* and *c*. Cannot be easily linearised (1 mark)

(c) Non-linear with respect to parameter *c*, but linear with respect to the other parameters. Cannot be easily linearised because of term *dx* (1 mark)

(d) Non-linear with respect to parameter *w*, but linear with respect to the other parameters. Could be linearised by fixing the value of parameter *w*, the wavelength (1 mark)

(iii) The analysis shows that the logistic curve describes the observed Expression response over time well, with the Expression response increasing smoothly from an initial low level, up to about Day 23, to reach a higher level by about Day 33 (1 mark). The initial level is given by parameter *a* at 8.839 units (1 mark) with the total increase in Expression given by parameter *c* at 2.617 units, so that the final Expression level is given by *a*+*c* at 11.456 units (1 mark). Parameter *m* identifies the time (about half-way between Day 27 and Day 28) at which the Expression has reached half-way between the lower asymptote (*a*) and upper asymptote (*a*+*c*) (1 mark). Parameter *b* is related to the maximum rate of change of the Expression level, achieved at time *m*, with larger values of *b* associated with a more rapid rate of change (1 mark). To estimate the time at which a gene expression response of 10.5 is reached, the equation needs to be inverted, so that *x* is expressed as a function of *y*. This is given by (1 mark):

$$x = m - \frac{1}{b}\ln\left(\frac{c}{y-a} - 1\right)$$

Substituting the fitted parameter values and y = 10.5 gives x = 28.59 days (1 mark).

(iv) Since there were replicate measures of the gene expression level at each time point, the analysis could be modified to split the residual variation into lack-of-fit (relating to the differences between the fitted curve and the mean of the observed values at each time point) (1 mark) and pure error (relating to differences between the individual replicate values and the mean observed value at each time point) (1 mark). The ratio of the lack-of-fit mean square to the pure error mean square then provides a formal test of the lack-of-fit of the model to the data (1 mark).

Question 5

(i) There are significant main effects of all three treatment factors, duration, manufacturer and temperature (1 mark), as well as significant interactions between duration and manufacturer, and between manufacturer and temperature (1 mark). Components produced by manufacturer B have a significantly higher failure rate than the other two manufacturers, with failure rates also increasing significantly with both test duration and the temperature at which the test was conducted (1 mark). The significant duration-by-manufacturer interaction indicates a greater increase in failure rate with increasing duration for components made by manufacturer B than for the other manufacturers, particularly between durations of 2 and 5 hours (1 mark). The significant manufacturer-by-temperature interaction indicates a greater increase in failure rate with increasing temperature for components made by manufacturer B than for the other manufacturers, between temperatures of 0 and 10 (1 mark). There is no evidence of differences in failure rates between the three independent laboratories (1 mark). The non-significant interaction between duration and temperature indicates that while failure rate increases as both increase, the increase with temperature is similar at all durations (and vice versa) (1 mark).

(ii) The assumptions associated with the analysis are that the residuals are distributed following a Normal distribution with mean zero (0.5 marks), that the variance of the residuals is constant (does not vary with the mean response – homogeneity of variances) (0.5 marks), and that the residuals are independent (0.5 marks). The residuals are estiamted as the difference between each observation and the value given by the fitted model (0.5 marks). The histogram of these estimated residuals can be used to assess the assumption of Normality (0.5 marks). The Fitted value plot is created by plotting each estimated residual against the value given by the fitted model, and allows assessment of the assumption of homogeneity of variances by assessing whether the spread of the estimated residuals changes as the fitted values increase (1 mark). The Normal and Half-Normal plots are obtained by ordering the estimated residuals or absolute values of the estimated residuals, and plotting these against the Normal quantiles, obtained as the mid-points of equi-probably intervals from a Normal distribution with the same number of observations (1 mark). A linear relationship between the estimated residuals and the Normal quantiles indicates that the estimated residuals follow a Normal distribution (0.5 marks).

(iii) The data are percentage failure rates, calculated as the number of failures from a fixed number of components tested (1 mark). The number of failed components might therefore be expected to follow a Binomial distribution for each treatment combination (1 mark), with the consequence that the variance of the residuals would be expected to vary with the mean response (fitted value) (1 mark) with smaller variances for treatment combinations with a mean failure rate close to 0% or 100%, and larger variances for treatment combinations with a mean failure rate close to 50% (1 mark). This pattern is visible to some extent to the right side of the Fitted value plot, where the variance appears to decrease with increasing fitted values, though is less clear at the left side of this plot (1 mark).

(iv) The simplest approach to overcome this issue is to apply a transformation to the percentage failure rate data prior to analysis (0.5 marks). Appropriate transformations include the arcsin (or angular) transformation, the logit transformation and the probit transformation (0.5 marks). An alternative analysis approach is to analyse the data within the generalized linear model framework (1 mark), assuming a Binomial distribution for the errors (0.5 marks) with either a logit or probit link function used to connect the expected values to the linear model of explanatory factors (0.5 marks).

Question 6

(i) In a confounded factorial design all factorial treatment combinations are included (1 mark), but with some, usually high order, interaction terms confounded or confused with blocks (0.5 marks), and therefore un-estimable (0.5 marks). In a fractional factorial design, a fraction of the complete set of factorial treatment combinations is selected to provide information about the main effects and low-order interactions of most interest (1 mark), with high-order interaction terms assumed to be negligible and assigned to the error (1 mark). The two approaches might be combined when resources only allow a fraction of the full set of factorial treatment combinations to be studied (0.5 marks), and where block size is still smaller that the size of the selected fraction (0.5 marks).

(ii) For a half-replicate the obvious defining (fractionating) contrast to use is ABCDEF (1 mark). This means that each main effect is aliased with a five-factor interaction, each two-factor interaction is aliased with a four-factor interaction, and pairs of three-factor interactions are also aliased (1 mark). If we want to obtain information about all of the main effects and most of the two-factor interactions (with the exception of EF, the interaction between the two temperature effects), then we want to choose two non-aliased three-factor interactions as confounding contrasts (1 mark). These should be chosen so that their generalised interaction is the EF term that we are not interested in (or, equivalently, the ABCD interaction with which this term is aliased) (1 mark). So, for example, choosing ABE and CDE will achieve this (1 mark).

(iii) The principal half-fraction for fractionating contrast ABCDEF will contain all factorial treatment combinations with an even number of factors at the upper level (1 mark): [1], ab, ac, ad, ae, af, bc, bd, be, bf, cd, ce, cf, de, df, ef, abcd, abce, abcf, abde, abdf, abef, acde, acdf, acef, adef, bcde, bcdf, bcef, bdef, cdef and abcdef (2 marks). The 8 treatment combinations in each block are then identified based on whether they have odd or even numbers of factors in the confounding contrasts (1 mark). For confounding contrasts ABE and CDE, the four blocks are (2 marks):

Block I:	ac, ad, bc, bd, ef, abef, cdef, abcdef
Block II:	af, bf, ce, de, abce, abde, acdf, bcdf
Block III:	ae, be, cf, df, abcf, abdf, acde, bcde
Block IV:	[1], ab, cd, abcd, acef, adef, bcef, bdef

(iv) The outline analysis of variance table is as follows (marks for identifying the confounding contrasts (1), the main effects (1) and the estimable two-factor interactions (1)):

SOURCE	DF	
Block stratum		
EF ABE / CDF CDE / ABF	1 1 1	Confounding contrasts
Block.Plot stratum		
Main effects (A to F)	6	
2-factor interactions	14	Not EF (confounding contrast)
Residual	8	Three-factor interactions
Total	31	

The assumption that is needed to allow assessment of the significance of the main effects and two-factor interactions is that the three-factor and higher order interactions are negligible. This allows the eight remaining degrees of freedom associated with three-factor interactions to be assigned to the residual, and the observed variation to be assigned to the main effects and two-factor interactions (excluding EF) rather than the higher order terms with which they are aliased. With only 8 residual degrees of freedom the experiment might still be considered under-powered, but we are interested in all 2-factor interactions other than EF (1 mark).

Question 7

(i) For fixed effects interest is in the parameters associated with each level of the fixed effect, which are constants to be estimated, and inferences only apply to the actual set of levels included in the experiment (1 mark). Fixed effects are usually used for treatment terms in a designed experiment, where interest is in the effect of each treatment, or, more usually, the differences between treatments (1 mark). For random effects interest is in the variability associated with the possible levels of the random effect, with the parameters being random variables, and the particular levels used in the experiment being a random sample from a much wider population of levels which could have been used (0.5 marks). Inferences are assumed to be valid for this wider population (1 mark). The blocks in a designed experiment will often be assumed to be random effects, though a better example would be a random sample of hospitals selected from a larger population to be used in a medical study (0.5 marks).

(ii) y_{ijk} is the water quality value measured for the sample collected on the k^{th} date (k = 1, 2, 3, 4, 5) at the j^{th} location (j = 1, 2, 3) for the i^{th} type of water treatment plant (i = 1, 2, 3, 4) (1 mark).

μ is a fixed term representing the overall grand mean water quality (1 mark).

Since the water company is interested in assessing the efficacy of the four particular types of water treatment plant, with no suggestion that these are a sample from a wider population of water treatment plant types, t_i is a fixed term representing water treatment plant type effects, with $\Sigma t_i = 0$ (1 mark).

The sampling locations have been selected at random from those available for each water treatment plant type, so l_{ij} is a random term representing variation between locations for water treatment plant type i (1 mark). It is assumed that the l_{ij} are Normally distributed with mean zero and variance σ_l^2. Note that σ_l^2 is assumed to be constant over i, i.e. the underlying variance is the same for the different types of water treatment plants (1 mark).

ε_{ijk} is a random residual (error) term, distributed following a Normal distribution with mean zero and variance σ^2 for all observations, i.e. the underlying variance is assumed to be the same for the different sampling locations (1 mark).

(iii) There are 60 samples collected (4 types, 3 locations for each type, 5 samples at each location), so 59 degrees of freedom in total. With 4 types of water treatment plant, there are 3 degrees of freedom for differences between water treatment plant types (0.5 marks). For each water treatment plant type, there are 3 locations, giving 2 degrees of freedom per type and there are therefore 8 degrees of freedom altogether for variation between locations within types (0.5 marks). This leaves 48 degrees of freedom for variation between samples, which is obtained as 4 degrees of freedom (between 5 sample dates) for each of the 12 locations (3 locations for each of 4 water treatment plant types) (1 mark).

The completed analysis of variance is therefore (1 mark for mean squares, 1 mark for F-test statistics):

Source of variation	d.f.	Sum of squares	Mean squares	*F* value
Between types	3	2562.82	854.27	854.27/50.50 = 16.92
Between locations within types	8	404.00	50.50	50.50/4.34 = 11.64
Between samples within locations	48	208.31	4.34	
Total	59	3175.13		

To test the null hypothesis "all $t_i = 0$", we compare 16.92 with the *F* distribution with 3 and 8 degrees of freedom (0.5 marks). This is significant at even the 0.1% significance level (the upper 0.1% point is 15.83), so there is strong evidence against this null hypothesis. We should conclude that there are important differences in water quality between water treatment plant types (1 mark).

To test the null hypothesis that $\sigma_l^2 = 0$, we compare 11.64 to the *F* distribution on 8 and 48 degrees of freedom (0.5 marks). This is very highly significant (the upper 0.1% point is 4.03) so we reject this null hypothesis, and conclude that there is very strong evidence that there is variability between locations within water treatment plant type (1 mark).

The variance between samples (dates) is estimated by 4.34 (1 mark).

Using the formulae for the expected mean squares given in the question, the variance between locations, σ_l^2, is estimated by (50.50 - 4.34)/5 = 9.23. This is just over twice the estimated variance between samples (dates) (1 mark).

In summary there are clearly differences in water quality associated with the different water treatment plant types, with variation between locations also being important (1 mark).

Question 8

(i) A generalized linear model assumes that responses (Y) come from a distribution from the exponential family (1 mark), with the expected value of the response variable connected to the linear model for the effects of any explanatory factors or variables through a link function (1 mark), $E[Y] = g^{-1}(X)$, where g() is the link function and X is the explanatory model (one or more factors and variates) (1 mark). For count data, as presented in this questions, the appropriate distribution function is the Poisson distribution (1 mark), with the logarithmic link function (the canonical link function) providing the connection to the linear model (1 mark).

(ii) Fitting a GLM involves the regression of an adjusted dependent variable, s (obtained using a linearised form of the link function applied to the dependent variable, y) using weights that are functions of the fitted values, against the explanatory variables (1 mark). The process is iterative because both the adjusted dependent variable, z, and the weights, w, depend on the fitted values, for which only current estimates are available (1 mark). The iterative procedure involves forming the adjusted dependent variable and weights based on the current estimate of the linear predictor (0.5 marks), regressing this adjusted dependent variable on the explanatory variables to obtain new estimates of the parameters (0.5 marks), using these new parameter estimates to form a new estimate of the linear predictor (0.5 marks) and repeating this process until changes between iterations are small (0.5 marks). Other equivalent descriptions should be accepted and marked according to their quality.

(iii) The accumulated analysis of deviance indicates a strong (significant) overall impact of tachyporus on aphid numbers but no apparent impact of melanarius (1 mark). There was no apparent interaction between the two predator species, suggesting neither synergistic nor antagonistic behaviour (1 mark). There was no apparent impact on aphid numbers of the fungicide application to control mildew, but the efficacy of the tachyporus predator was influenced by the application of the fungicide (1 mark). There was considerable variation in the number of aphids between periods, and also of the efficacy of the tachyporus predator, and combined efficacy of the two predators, between periods. (1 mark). The efficacy of the different treatments can be obtained from the "antilog of estimate" value given for the different parameters (1 mark). In the absence of the other predators or fungicide treatment, the efficacy of the tachyporus predator is to significantly reduce aphid numbers by about 68% (obtained as (1 - 0.3198)*100) (0.5 marks), while the melanarius predator appears to be ineffective alone, resulting in a non-significant increase of about 11% in aphid numbers (obtained as 1 - 1.114)*100) (0.5 marks). The combined efficacy of the two predators is obtained by adding together the estimates for the separate effect of the tachyporus predator (-1.140), the separate effect of the melanarius predator (0.108) and the combined effect (0.425), giving an overall effect of -0.607 relative to the response in the absence of both predators (1 mark). The combined efficacy is then obtained from taking the exponential of this value, giving a reduction in aphid numbers of about 46% (obtained as (1 - exp(-0.607)*100) (0.5 marks). The effect of the fungicide application is a non-significant reduction in aphid numbers of about 37% (obtained as (1 - 0.6341)*100) (0.5 marks).

(iv) Overdispersion indicates that there is more variability (1 mark) than would be expected under a Poisson distribution, indicating, in this case, that the occurrence of individual aphids on plants are not independent events, but that the underlying distribution is for aggregated counts (1 mark). Evidence for overdispersion in the output is that the residual mean deviance is substantially greater than 1.00 (1 mark) (as would be the case under a pure Poisson distribution).