HONG KONG STATISTICAL SOCIETY

2015 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) A balanced incomplete block design has a block size smaller than the number of treatments, so that not all treatments can be included in each block (1). Treatments are allocated to blocks so that all pairwise occurrences of treatments in blocks occur equally often across the whole design, with treatments having equal replication (1). Balanced incomplete block designs exist where the five parameters included in the relationships all have integer values (1). The first relationship is concerned with the total number of observations in the design, given either by the number of treatments multiplied by the number of replicates, or by the number of blocks multiplied by the block size (1). The second relationship is concerned with the number of pairwise comparisons for a particular treatment, given either by the number of times a pair of treatments occur together in blocks multiplied by the number of treatments to be compared, or by the number of replicates per treatment multiplied by the number of comparisons possible in each block in which the treatment occurs (1). Given values of t and λ, the second relationship can first be used to find possible integer combinations of r and k given the maximum possible block size, and given these values also, the first relationship can then be used to identify whether the number of blocks is also an integer value (1).

(ii) With 8 treatments, the block size (k) can be no larger than 7 for a BIBD (1).

a. For $\lambda = 1$, $r(k\text{-}1)$ must equal 7, which is only possible with $r = 7$ and $k = 2$ (as k cannot be 8). With these values we have $2b = 8*7$, so $b = 28$ (1).

b. For $\lambda = 3$, $r(k\text{-}1)$ must equal 21, which is only possible with $r = 7$ and $k = 4$ (as k cannot be 8). With these values we have $4b = 8*7$, so $b = 14$ (1).

c. For $\lambda = 6$, $r(k\text{-}1)$ must equal 42, which is only possible with $r = 7$ and $k = 7$ (as k cannot be 8). With these values we have $7b = 8*7$, so $b = 8$ (1).

For $\lambda = 2$, $r(k\text{-}1)$ must equal 14, which is only possible with $r = 7$ and $k = 3$ (as k cannot be 8). With these values we have $3b = 8*7$, for which there is no integer solution (1).

(iii) The design with a block size of 2 ($\lambda = 1$) may be appropriate for particular applications where this is the natural block size, but in general is not a good choice as much of the information about differences between treatments is based on between-block comparisons, requiring a combined analysis from the within- and between-block comparisons to provide a good estimate of treatment differences (1). In contrast, the design with a block size of 7 ($\lambda = 6$) is almost as good as the randomized complete block design, but often the natural block size will either be smaller than this, or there will be sufficient flexibility to allow the randomized complete block design to be used (1). Finally, the design with a block size of 4 ($\lambda = 3$) provides a sensible compromise, with most of the information about differences between treatments based on within-block comparisons so that the within-block analysis is probably sufficient, but a block size that is likely to be within the range of natural block sizes and therefore easily achievable (1).

(iv) Given this constraint, the design with a block size of 4 ($\lambda = 3$) should be constructed. This design (with 14 blocks) can be constructed by sequentially adding treatments to blocks to meet the conditions of each treatment occurring in 7 different blocks and any pair of treatments only occurring together in 3 blocks (1). Treatment A is allocated to the first 7 blocks, with treatment B added to 3 of these and to the next 4 blocks (1). Treatment C appears in the first block with treatments A and B, then in two further blocks with A only, two further blocks with B only, and in two further blocks (1). Care is needed with the allocation of treatment D – this can appear in the first blocks with treatments A, B and C, filling the block, but must then appear in the six blocks with only one treatment already allocated, so twice with A only, twice with B only and twice with C only (1). For the remaining treatments we note that the final block still has four empty spaces and so we allocate all 4 remaining treatments (E, F, G and H), then note that the 12 blocks with empty spaces (two each) form 6 pairs of blocks, with the blocks in each pair having the same treatments already allocated (AB, AC, AD, BC, BD or CD) (1). The 4 remaining treatments must be allocated once to each of these 6 pairs of blocks, so that each pair of treatments (EF, EG, EH, FG, FH and GH) occur exactly twice. There are various ways of achieving this, but the simplest approach has the same pairs of treatments (from E, F, G and H) allocated to sets of pairs of blocks with complementary pairs of treatments (from A, B, C and D). So the pairs of blocks (2/3 and 12/13) with AB and with CD get the same allocation of pairs of treatments from E, F, G and H, as do the pairs of blocks (4/5 and 10/11) with AC and BD, and the pairs of blocks (6/7 and 8/9) with AD and BC (1). The allocation is:

Block	1	2	3	4	5	6	7	8	9	10	11	12	13	14
	A	A	A	A	A	A	A	B	B	B	B	C	C	E
	B	B	B	C	C	D	D	C	C	D	D	D	D	F
	C	E	G	E	F	E	F	E	F	E	F	E	G	G
	D	F	H	G	H	H	G	H	G	G	H	F	H	H

Question 2

(i) The key issue to consider here is the possibility of variation in pest pressure in two orthogonal directions across the array of 6 rows and 6 columns (though alternative plot shapes could be used to provide a linear array of plots, or fewer plots in one of the orthogonal directions).

a. Use of a completely randomized design in this scenario would be assuming that the 36 field plots available for the trial would behave identically if the same treatment was applied to each (1). However, it is expected that there will be variation in pest pressure across the trial, so that a completely randomized design might result in a randomisation such that all six replicates of one treatment were at one side of the trial and hence subjected to the highest pest pressure (1). The dummy analysis of variance table is (1):

Source	Degrees of freedom
Seed treatment	1
Foliar spray	2
Interaction	2
Residual	30
Total	35

A possible layout might look like this with no pattern to the arrangement of treatments, and replicates of some treatments being grouped together (1):

A	A	C	B	E	B
C	A	E	F	D	E
B	F	F	C	E	A
D	E	D	D	F	C
C	A	F	A	B	D
D	B	E	C	F	B

b. Use of a randomized complete block design allows the blocking of the plots in one of the two orthogonal directions, which might be okay if this happens to coincide with the direction from which the insect pests arrive (1). But if the major direction in pest pressure is orthogonal to blocks, this design will be as poor as the completely randomized design, with the possibility that all six replicates of one treatment might be at one side of the trial and hence subjected to the highest pest pressure (1). The dummy analysis of variance table is (1):

Source	Degrees of freedom
Block stratum	5
Block.Plot stratum	
Seed treatment	1
Foliar spray	2
Interaction	2
Residual	25
Total	35

A possible layout might look like this with blocks as rows (could be blocks as columns) and each treatment appearing once in each block (1):

A	B	C	D	E	F
B	A	C	F	E	D
C	A	E	F	D	B
E	F	C	A	B	D
A	C	E	D	F	B
F	A	B	D	C	E

Or this (with more compact blocks which is better) (1 – in addition to the above mark for the sketch):

A	B	C	D	A	B
D	E	F	E	F	C
A	C	D	A	C	E
B	E	F	B	F	D
B	C	E	C	E	D
F	A	D	A	B	F

c. Use of a Latin square design provides the best approach to cope with the unknown direction from which the pests will arrive, with each of the 6 treatments occurring once in each column and once in each row (1), and so each treatment should experience the full range of pest pressures no matter from which direction the pest arrives (1). The dummy analysis of variance table is (1):

Source	Degrees of freedom
Row stratum	5
Column stratum	5
Row.Column stratum	
Seed treatment	1
Foliar spray	2
Interaction	2
Residual	20
Total	35

A possible layout might look like this, with each treatment appearing exactly once in each row and once in each column (1):

A	B	C	D	E	F
B	C	D	E	F	A
C	D	E	F	A	B
D	E	F	A	B	C
E	F	A	B	C	D
F	A	B	C	D	E

(ii) Randomization of treatments for a Latin square design should first identify the set of standard Latin squares from which a random square will be selected (1). Once a square has been randomly selected, the rows and columns of the square should be separately randomised (1). Finally, the six treatment combinations should be randomly allocated to the six treatment codes (A, B, C, D, E and F) used to identify the different treatments in the Latin square (1).

(iii) (a) This design approach is usually referred to as a split-plot or split-unit design (0.5). It is most sensible to group together adjacent plots either within rows (or within columns), so that a complete row (or a complete column) will contain three main plots (pairs of plots), and a complete column (or a complete row) will contain six main plots (0.5). It would be sensible to adopt a layout borrowing ideas from the RCBD or (preferably) the Latin Square design, so that each row (or column) is a block (and preferably each column is also a block – with two replicates of the spray treatments) – see diagram below (1). There are two steps to the randomisation process, firstly of spray treatments to the pairs of plots (taking account of any other structure imposed) then of seed treatments to the plots within each pair separately (1)

(b) The main disadvantage of using such a design is that the effects of the seed treatments and spray treatments are estimated with different levels of precision (0.5), with relatively few degrees of freedom for assessing the main effect of the spray treatments (0.5).

A	D	E	B	C	F
F	C	D	A	E	B
B	E	C	F	A	D
E	B	A	D	C	F
C	F	E	B	D	A
D	A	F	C	B	E

Question 3

(i) In a confounded factorial experiment all factorial treatment combinations are included (1), but with some, usually high-order, interaction terms confounded or confused with blocks (1), and therefore un-estimable. 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), with high-order interaction terms again assumed to be negligible and assigned to the error (1). Fractionating contrasts are used to define the particular fraction of factorial treatment combinations to be included, and are selected by considering the different model terms that will be aliased (1). Confounding contrasts are used to split the factorial treatment combinations into blocks, identifying the model terms that will be confounded with block effects (1).

(ii) Two fractionating contrasts are needed to specify a quarter fraction, with a third fractionating contrast defined by their generalised interaction (0.5). As interest is in the main effects and two-factor interactions, fractionating contrasts need to be chosen so that these terms are not aliased (0.5). Two carefully chosen five-factor interaction terms will generate a four-factor generalised interaction term, ensuring that main effects are only aliased with three-factor or higher-order interactions (0.5), though three pairs of two-factor interactions will be aliased (0.5). Example pairs of five-factor interaction terms to use, and the generalised four-factor interaction term generated, include (ABCDE, ACEFG => BDFG), (ABDEF, ACDFG => BCEG), (ABCFG, BDEFG => ACDE) (1 – for an appropriate choice). Choice from these (and others) will depend on any information on the two-factor interactions of most interest.

(iii) For the first set of fractionating contrasts (ABCDE, ACEFG => BDFG), the main effects are aliased as follows (3 marks for a complete correct set of aliased terms, with a reduction for any errors):

A ≡ BCDE ≡ ABDFG ≡ CEFG
B ≡ ACDE ≡ DFG ≡ ABCEFG
C ≡ ABDE ≡ BCDFG ≡ AEFG
D ≡ ABCE ≡ BFG ≡ ACDEFG
E ≡ ABCD ≡ BDEFG ≡ ACFG
F ≡ ABCDEF ≡ BDG ≡ ACEG
G ≡ ABCDEG ≡ BDF ≡ ACEF

(iv) The principal quarter fraction is the fraction that includes the factorial treatment combination with the lower level of each of the factors, usually denoted [1]. The factorial treatment combinations that are also in this fraction are those that include an even number of upper levels of factors that appear in each of the fractionating contrasts (1). So, for the first set of fractionating contrasts (ABCDE, ACEFG -> BDFG) the principal quarter fraction is: (3 marks for a compete correct set of combinations – lose 1 mark for the wrong fraction, with a reduction for any errors), ac, ae, bd, ce, fg, abf, abg, adf, adg, bcf, bcg, bef, beg, cdf, cdg, def, deg, abcd, abde, acfg, aefg, bcde, bdfg, cefg, abcef, abceg, acdef, acdeg, abcdfg, abdefg, bcdefg.

(v) To identify two confounding contrasts to divide the quarter fraction into four blocks each containing 8 treatment combinations, we choose terms that are not aliased with any of the main effects, remembering that the generalised interaction will also be confounded with blocks (1). A possible set for the above aliasing scheme is ABE and ACF, with generalised interaction BCEF (1). This generates the following division of the above principal fraction, based on whether the treatment combination has an odd or even number of “letters” included in each of the two confounding contrasts (2 marks for a completely correct division, with reductions for errors):

Block 1: [1], abf, beg, cdf, abcd, aefg, acdeg, bcdefg
Block 2: ac, bd, adf, bcf, deg, cefg, abceg, abdefg
Block 3: ae, fg, abg, bef, cdg, bcde, acdef, abcdfg
Block 4: ce, adg, bcg, def, abde, acfg, bdfg, abcef

We can check that each of the main effects is still estimable by noting that each letter appears in exactly four of the terms in each block. This particular choice also confounds the CD and EG interaction terms with blocks (aliased with ABE and ACF respectively), but we were not required to ensure that all 2-factor interactions were still estimable.

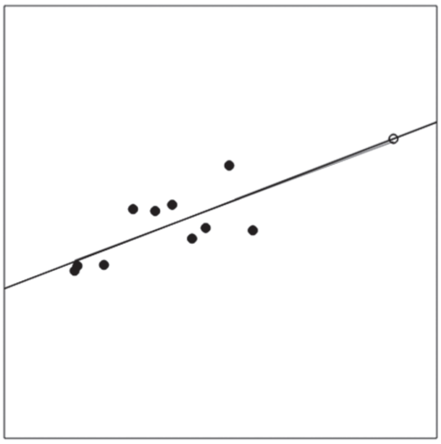
Question 4

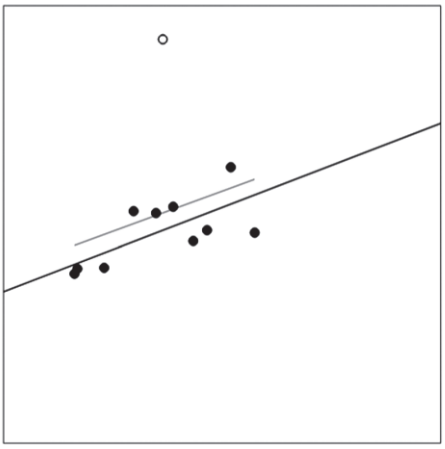
(i) Leverage is a measure that identifies the more extreme values of an explanatory variable, which have the potential to be highly influential on the model fit (0.5). However, leverage cannot quantify whether the observation has had a large impact on the fitted model. The influence of an observation is a measure of the change in the fitted values that would occur if that observation was omitted (0.5). The leverage values are the diagonal elements of the hat matrix ($\boldsymbol{X}\,(\boldsymbol{X}^T\boldsymbol{X})^{-1}\,\boldsymbol{X}^T$) (1). Candidates might alternatively or additionally note that in the case of simple linear regression these can be calculated directly as:

$$h_{ii} = \frac{1}{N} + \frac{(x_i - \bar{x})^2}{SS_x}$$

where N is the sample size, and SS_x is the sum of squares of the explanatory variable, *x*.

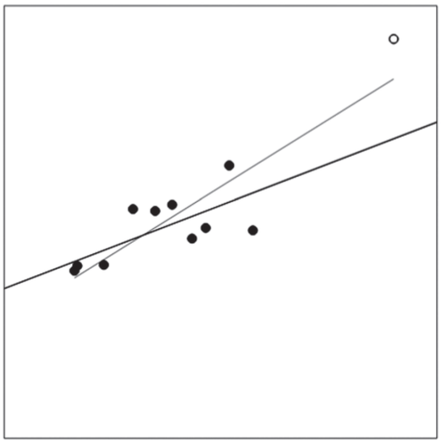
(a) the open circle in the first plot (below, left) has high leverage and low influence – the point is extreme within the set of values of the explanatory variable, but lies on the regression line obtained for the remaining points (1 – for plot and description).





(b) the open circle in the second plot (right, above) has low leverage but high influence – the point is in the centre of the range of explanatory variable values, but the inclusion of this point shifts the regression line up relative to the other points (1 – for plot and description).

(c) the open circle in the third plot (below) has high leverage and high influence – it is both extreme within the set of explanatory variable values and inclusion of this point dramatically changes the slope of the regression line (1 – for plot and description).



Variables X2 and X3 both have one extreme large value (which, from the plots of Y against each variable can be identified as being associated with the same observation) (0.5). The other points with large leverages are most likely to be associated with X1 (one larger value) and X6 (one smaller value), as these points are the most extreme from the distributions for each of the explanatory variables (0.5) – other arguments about potential extreme values of the explanatory variables would also be acceptable.

(ii) The coefficient of determination measures the proportion of variation in a dataset that is accounted for by the fitted model, calculated as the ratio of the model sum of squares to the total sum of squares, or alternatively as one minus the ratio of the residual sum of squares to the total sum of squares (1). The adjusted coefficient of determination also takes into consideration the sample size and the number of estimated parameters, and is calculated as one minus the ratio of the residual mean square to the total mean square (1). The coefficient of determination will never decrease as additional explanatory variables are added to the model, because the residual sum of squares will never increase when adding new explanatory variables (1). The adjusted coefficient of determination will only increase if the additional variable decreases the residual sum of squares by a proportion greater than the proportional change in the residual degrees of freedom – so if the adjusted coefficient of determination does not increase this provides evidence that the additional variable is not improving the model (1).

(iii) The Mallows C_p statistic correspond to a situation with a total of m potential explanatory variables, and is used to compare the fit of a model containing q of these variables to the full model containing all of the m variables (1). The value of the statistic for the full model always equals m+1, the total number of parameters in that model (1). Any sub-model containing q variables that has a similar value of the residual mean square (and hence similar precision) to the full model will have a C_p value close to p (= q+1), the number of fitted parameters for that sub-model (1). Plotting values of C_p against p for competing models, along with the line $C_p = p$, provides a simple screening process – good models will have points close to this 1:1 line (1). Plotting the C_p values against the Df values from the presented results, the best 5-term model is closest to the line, with the best 4-term model not much further away – so either of these two models would be appropriate to consider (1).

(iv) The most important variable appears to be the number of manufacturing enterprises employing 20 or more workers (X2), with levels of sulphur dioxide increasing as this variable increases (1). Whilst population size (X3) appears to have a positive association from the scatter plots (0.5), the estimate for this variable is negative, indicating that as population increases the level of air pollution decreases, almost certainly a compensatory effect given the strong correlation between variables X2 and X3 (1) – so the increase in air pollution with number of manufacturing enterprises does not increase as fast when the population size is also increasing (0.5). Levels of air pollution also seem to decrease with increasing temperature (X1), presumably because this reduces the need for heating (0.5), with a decrease in air pollution also as the average wind speed (X4) increases, presumably because this disperses the pollutants (0.5). Note that average wind speed and average temperature appear to be slightly negatively correlated, so that these last two effects would seem to be competing (0.5). The full model suggests that levels of sulphur dioxide are not influenced by rainfall, expressed as either the average annual rainfall (X5) or the number of days with rainfall per year (X6) (0.5).

Question 5

(i) The correction factor is $(163.23)^2/40$ = 666.101, so the total sum of squares is 736.784 – 666.101 = 70.683 (0.5)

The Block SS is $\frac{41.28^2}{10}+\frac{42.06^2}{10}+\frac{42.30^2}{10}+\frac{37.59^2}{10}-666.101=667.538-666.101=1.437$ (0.5)

The Variety SS is $\frac{66.21^2}{20}+\frac{97.02^2}{20}-666.101=689.832-666.101=23.731$(1)

The Phosphorous SS is $\frac{20.68^2}{8}+\frac{31.64^2}{8}+\frac{34.88^2}{8}+\frac{35.42^2}{8}+\frac{40.61^2}{8}-666.101=693.639-666.101=27.538$ (1)

The Variety-Phospohrous interaction SS is $\frac{9.41^2}{4}+\frac{13.82^2}{4}+\cdots+\frac{24.52^2}{4}-23.731-27.538-666.101$ = 722.141 – 23.731 – 27.538 – 666.101 = 4.772 (1)

Hence

SOURCE	DF	SS	MS	F-value	
Blocks	3	1.437	0.479	0.982	non-significant
Variety	1	23.731	23.731	48.629	very highly significant
Phosphorous	4	27.538	6.885	14.109	very highly significant
Variety.Phosphorous	4	4.772	1.193	2.445	borderline
Residual	27	13.205	0.489		
TOTAL	39	70.683			

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

Key elements to comment on are the large effect of variety (0.5) and the large variation associated with phosphorous levels (0.5). It would certainly be sensible to explore the shape of the response to phosphorous, as in the rest of the question, and also the consistency of this effect across varieties (given the borderline F-value of the interaction term) (0.5). There is no evidence for differences between blocks, but as the experiment was designed as a RCBD this term should be retained in this analysis (though combining it with the residual would have very little impact) (0.5).

(ii) Orthogonal contrasts lead to independent terms in an analysis of variance, and thus to independent tests and comparisons (1). 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 variance, and, this, by testing each one against the residual in the usual way, gives a mechanism for investigating where any treatment differences lie (or, in the case of polynomial contrasts, indicate the shape of the response) (1). The orthogonality of a pair of contrasts can be tested by calculating the sum of the products of the coefficients for the contrasts – if this sum is zero then the contrasts are orthogonal (1).

(iii) As the sum of the coefficients of a contrast must equal zero, the simplest integer coefficients for a linear contrast for a five-level factor are (-2, -1, 0, 1, 2) (1). A simple way of obtaining the coefficients for the quadratic contrast is to square these values and then adjust so that they sum to zero. This gives the coefficients (2, -1, -2, -1, 2) (1). The sum of the products of the coefficients is (-2*2 + -1*-1 + 0*-2 + 1*-1 + 2*2) – the first and last terms cancel out, as do the second and fourth, leaving a sum of zero, and hence these two contrasts are orthogonal (1).

(iv) The linear contrast value is $\frac{(-2)*20.68+(-1)*31.64+(0)*34.88+(+1)*35.42+(+2)*40.61}{8}=\frac{43.64}{8}=5.455$ (0.5), so the sum of squares for this contrast is $\frac{43.64^2}{[(-2)^2+(-1)^2+(0)^2+(+1)^2+(+2)^2]*8}=23.806$ (0.5).

The quadratic contrast value is $\frac{(+2)*20.68+(-1)*31.64+(-2)*34.88+(-1)*35.42+(+2)*40.61}{8} = \frac{-14.24}{8} = -1.78$ (0.5), so the sum of squares for this contrast is $\frac{(-14.24)^2}{[(+2)^2+(-1)^2+(-2)^2+(-1)^2+(+2)^2]*8} = 1.811$ (0.5).

The linear contrast has a variance ratio of 48.783, so is clearly very highly significant, indicating a strong increasing trend for yield as levels of phosphorous increase (1). The quadratic contrast has a variance ratio of 3.711 which is only borderline (in-)significant, the negative coefficient suggesting that the rate of increase of yield reduces as the level of phosphorous increases (1). The sum of squares for the deviations about this quadratic response is 1.921 on 2 degrees of freedom, giving a variance ratio of 1.968, which is not significant, suggesting that there is no lack of fit about the quadratic response (1). If we choose to just fit a linear response, then the sum of squares for the deviations is 3.732 on 3 degrees of freedom, giving a variance ratio of 2.549 which is only borderline (in-)significant, suggesting that it might be sensible to include a quadratic component in the fitted model (1).

Question 6

(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 (0.5). 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 (0.5). A function *f*(*y*) can often be found from this relation such that Var(*f*(*y*)) is constant (as shown in part (ii)). Analysis is then carried out in terms of *f*(*y*) not *y*: *f* is a transformation to stabilise variance (1).

The fitted-value plot for the untransformed count data shows a common pattern associated with non-constant variance, with the variance (scatter of the residuals) increasing as the fitted value increases. This indicates the need to apply a variance stabilising transformation prior to analysis (1).

(ii) A Taylor series expansion about μ is

$$h(y) = h(\mu) + (y-\mu)h'(\mu) + \tfrac{1}{2}(y-\mu)^2 h''(\mu) + \cdots \text{ (1)}$$

So

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

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

Similarly to second order

$$Var\big(h(Y)\big) = E[(h(Y) - E[h(Y)])^2] = \{h'(\mu)\}^2 E[(Y-\mu)^2] = \sigma_Y^2\{h'(\mu)\}^2 \quad \text{(1)}$$

If now σ = *f*(μ), we have Var(*h*(*Y*)) = $\{f(\mu)\ h'(\mu)\}^2$ which is constant if $\frac{dh(y)}{dy} \propto \frac{1}{f(y)}$ (1)

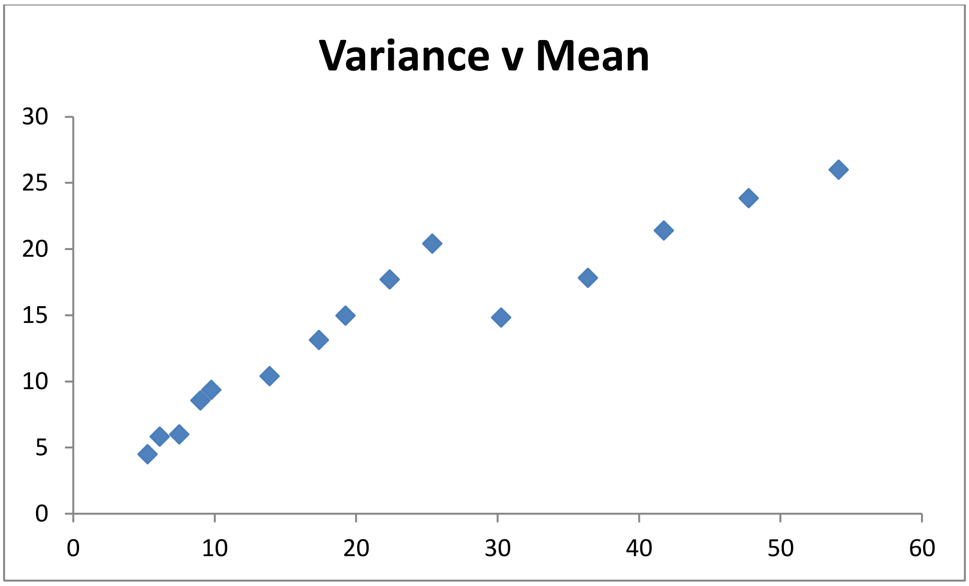
(iii) Noting that all transformations include a multiplicative constant

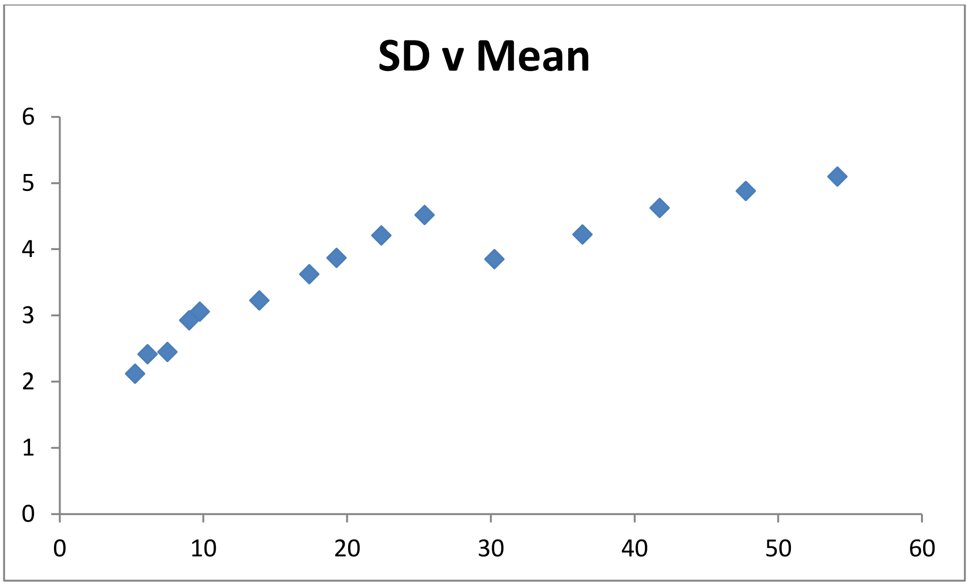
a. If the variance is proportional to the mean, then *f*(*y*) is equal to the square root of *y*, and $\int \frac{dy}{\sqrt{y}} = 2\sqrt{y}$, so the appropriate transformation to use is a square root. (1)

b. If the standard deviation is proportional to the mean, then *f*(*y*) = *y* and $\int \frac{dy}{y} = log(y)$, so the appropriate transformation is log *y*. (1)

(iv) While neither of the plots of the variance against the mean, and the standard deviation against the mean (produced by the candidate (1) and shown below) show a perfectly proportional relationship, the plot of the variance against the mean is certainly the straightest relationship (1), and so, of the two presented transformations, the square root transformation would appear to be the appropriate variance stabilising transformation for these data (0.5).

In the residual plots for the analysis of the square root transformed data, the residual pattern in the fitted value plot is certainly better than for the untransformed data (1), though there might still be some suggestion of the variance increasing with the mean, suggesting that a stronger transformation (such as the log transformation) might be needed (0.5).





(v) For the analysis of the untransformed data there is a significant species-by-concentration interaction in addition to the highly significant main effects (1). But in the analysis of the transformed data, the interaction is no longer significant, potentially simplifying the interpretation of the experiment (1). In the residual plots, the Normal plot for the transformed data shows a much straighter relationship than is seen for the untransformed data, suggesting that as well as dealing with the lack of homogeneity, the transformation has also improved the situation with regards to the assumption of Normality (1).

(vi) An alternative analysis approach for these count data would be to use a Generalised Linear Model analysis (1) assuming an underlying Poisson distribution for the data with a log link function (1). This would have the added benefit of being able to fit a regression relationship for the effect of concentration (1), allowing the assessment of whether the response to concentration is linear, and the assessment of whether the fitted slopes and intercepts are the same for the different species (1).

Question 7

(i) A generalised linear model assumes that responses (Y) come from a distribution from the exponential family (1), 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), $E[Y] = g^{-1}(X)$, where $g()$ is the link function and X is the explanatory model (one or more factors and variables) (1).

(ii) Fitting a GLM involves the regression of an adjusted dependent variable, z (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). 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). The iterative procedure involves forming the adjusted dependent variable and weights based on the current estimate of the linear predictor (0.5), regressing this adjusted dependent variable on the explanatory variables to obtain new estimates of the parameters (0.5), using these new parameter estimates to form a new estimate of the linear predictor (0.5), and repeating this process until changes between iterations are small (0.5). Other equivalent descriptions should be accepted and marked according to their quality.

(iii) The accumulated analysis of deviance table is:

Source	df	deviance	mean deviance	
Single line	1	916.3	916.30	Very highly significant
Separate intercepts	3	40.0	13.33	Highly significant
Separate slopes	3	0.7	0.24	Not significant
Residual	160	203.7	1.27	
Total	167	1160.7		

(2 marks for constructing the accumulated analysis of deviance table – 0.5 for structure, 0.5 for df, 0.5 for deviances, 0.5 for mean deviances)

Model 1 provides very strong evidence for an effect of insecticide concentration on the proportion of insects killed, averaged across all four insecticides (1). Comparing Model 2 with Model 1, giving the “Separate intercepts” line in the accumulated analysis of deviance, provides strong evidence for the need for a parallel line model to explain the different responses to the different insecticides, with separate intercepts for each insecticide but a common slope parameter (1). Comparing Model 3 with Model 2, giving the “Separate slopes” line in the accumulated analysis of deviance, indicates that there is no evidence for the need for different slope parameters for the different insecticides, indicating that the response to changes in insecticide log(dose) are similar for the different insecticides (1). So, Model 2 is the most appropriate model to use to describe the observed responses (1). The deviances should be compared with the appropriate critical values of a chi-square distribution to formally test the values (1). Alternatively, if it was considered that there was evidence for over-dispersion in the counts (the residual mean deviance is greater than 1.000 (0.5)), then approximate F-tests could be applied to the ratios of the model mean deviances to the residual mean deviance (0.5).

(iv) The standard parameterisation for a straight-line model is with a slope parameter and an intercept parameter, the intercept being the fitted value of the response when the value of the explanatory variable (in this case log(dose)) is zero. So, here we have

$$\text{logit}(p) = \text{intercept} + \text{slope}*\log_{10}(\text{dose}).\ (0.5)$$

To express this in terms of the LD50 instead of the intercept, we re-arrange the equation so that the intercept is obtained as minus the product of the slope and the LD50,

$$\text{logit}(p) = \text{slope}^*(\log_{10}(\text{dose}) – \text{LD50}).$$ (1)

So the LD50 is calculated as the intercept divided by the slope, with a change of sign. (0.5)
Because the fitted model for these data is a parallel lines model, the difference in efficacy between two insecticides does not change depending on the log(dose) that is considered (0.5). Equivalently, the difference in log(dose) for a particular level of efficacy will also be constant, whatever the level of efficacy chosen (0.5).
Based on the LD50 parameterisation of the model, the relative efficacy can best be represented as the change in LD50 values between insecticides (0.5) – that with the smaller LD50 is more effective, and, because the explanatory variable here was $\log_{10}$(dose), calculating 10 to the power of the difference gives the increase in efficacy of one insecticide over the other (0.5).
So for insecticide A and B, the LD50 values (on the $\log_{10}$ scale) are 1.440 and 1.659, so the difference in these is 0.219 (0.5). Calculating 10 to this power gives the relative increase in dose to achieve the same level of control, in this case the value is 1.656, so to achieve the same level of control as insecticide A we have to apply insecticide B at 1.656 times the dose applied for insecticide A (0.5).

Question 8

(i) The application of a GLM to the analysis of contingency table data assumes a multinomial distribution for the constrained counts (1). This can be modelled using the log-linear model framework, which is a GLM assuming a Poisson distribution and a log-link function (1), by considering the distribution of responses across the categories of one explanatory factor within the level of another explanatory factor as being conditional on the total count for that level (1).

(ii) The saturated model containing all the terms in the final model plus the four-factor interaction will have a deviance of zero with 0 degrees of freedom (1). This four-factor interaction essentially indicates that there are no consistent responses across the levels of the explanatory factors, but that the proportion of infections are different for every combination of Tap-water, Swimming and Hotel Catering (1). The change in deviance for adding this four-factor interaction term is 5.63 on 4 degrees of freedom, which is non-significant, indicating that a simpler, more parsimonious, model than the saturated model can be used to describe the data (1).

(iii) The Baseline model includes the main effects of the Respiratory Infection factor, plus all the main effects and interactions among the factors representing the possible causes of these infections (1). These terms identify the overall numbers of individuals with and without respiratory infections, and the numbers of individuals with each combination of the possible causal factors (1). Our interest in analysing the data is to determine whether there are any associations between different causal factors and the occurrence of respiratory infections (1). The terms in the baseline do not provide any information about these associations, but any more complex models will include terms that do (0.5). Hence we start from this baseline model explaining the marginal distributions of the counts, and check for associations that explain additional variability in the data (0.5).

(iv) Using a forward selection approach we start from the baseline model and add the most significant available term in at each step (where we cannot add a higher-order term in before all associated lower-0order terms have been included) (1). There are three possible models that can be obtained by adding one term to the baseline model, in each case adding an association between one of the three potential causal factors. All three changes in deviance are significant, but the most significant change (reducing the mean residual deviance the most) is to add in the interaction between RI and TW (Tap-water) (change of 34.57 on 1 df) (1). From this model there are two possible models that can be obtained by adding one term, adding an association with either of the other potential causal factors. Again, both changes in deviance are significant, but the most significant change is to add in the interaction between RI and SW (Swimming) (change of 37.23 on 2 df) (1). In the third step, there are again two possible models that can be obtained by adding one term to this model, either adding an association with Hotel Catering, or allowing the associations with Tap-water and Swimming to vary with the other factor (including the three-factor interaction of RI with TW and SW). Both changes are again significant, but the most significant change is the addition of the interaction between RI and HC (Hotel Catering) (change in deviance of 26.72 on 2 df) (1). Having included all three two-factor interactions of potential causal factors with the Respiratory infection factor, there are three possible models to consider by adding one further term, each involving the interaction between a pair of these causal factors and the Respiratory Infection factor. Only the addition of the interaction between TW, SW and RI is significant, and so this term is added to the model (change of 6.55 on 2 df) (1). No further additions of terms are significant, and so the best fitting and parsimonious model includes interactions between Respiratory Infection and the main effects of all three causal factors, plus the interaction between Tap-water and Swimming (1).

(v) This model indicates that levels of respiratory infection appear to vary between the different levels of each of the three causal factors (0.5), with different combinations of the Tap-water usage and Swimming patterns also changing the levels of respiratory infection (0.5). Examining the data table, it is clear that drinking Tap-water in the hotel is likely to be a causal factor, as is swimming in the Hotel Pool, with no clear association with swimming on the local beach (1). The interaction between these factors is because the levels of respiratory infections associated with swimming in the Hotel Pool are not particularly reduced for those not drinking Tap-water, suggesting that the Hotel Pool is the primary causal factor (1). In terms of Hotel Catering, there is an increased risk associated with eating in the Restaurant, with a further increase in risk with also drinking in the Bar (1).