ROYAL STATISTICAL SOCIETY

DATA | EVIDENCE | DECISIONS

THE ROYAL STATISTICAL SOCIETY

2015 EXAMINATIONS − SOLUTIONS

HIGHER CERTIFICATE – MODULE 6

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.

1. (i) A polynomial regression model of order 4 is

$$Y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4 + \epsilon_i$$

where Y_i is the random variable representing the i-th quality value, x_i is the corresponding i-th mixing speed (both in appropriate units). **(1)**
The random variation is represented by $\epsilon_i \sim N(0, \sigma^2)$, independent normal random variables with zero mean and constant variance σ^2 **(2)**
(1 mark for 3 of the 4 characteristics)

(ii) Consider two nested models where the 'full model', with one extra parameter, has a residual sum of squares RSS_F on df_F degrees of freedom and the 'reduced model' has residual sum of squares RSS_R. For backward elimination the F test formula is then

$$\frac{RSS_R - RSS_F}{RSS_F/df_F} \sim F_{1,df_F}$$

if the null hypothesis, that the extra parameter in the full model is zero, is correct. In polynomial regression we should try to omit the highest power of x at each stage. **(1)**
(This does not need to be stated explicitly as long as it is used correctly).
We first try to omit x^4 from the model, so the first test is of $H_0 : \beta_4 = 0$. The full model has $RSS_F = 231.3$ on $df_F = 54$ giving a mean square of $RSS_F/df_F = 4.2833$. Omitting x^4 the RSS increases from $RSS_F = 231.3$ to $RSS_R = 246.6$ using one degree of freedom, so that the test statistic is

$$F_{obs} = \frac{246.6 - 231.3}{4.2833} = 3.5720 \qquad (\mathbf{1})$$

The critical value is the upper 5% value of $F_{1,54}$ which is 4.0195 (students may need to interpolate or use the closest value in tables here). Hence $F_{obs} < F_{crit}$ so H_0 is not rejected and x^4 can be omitted. **(1)**
Next we try to omit x^3 from the model and so test $H_0 : \beta_3 = 0$. The full model now has $RSS_F = 246.6$ on $df_F = 55$ giving a mean square of $RSS_F/df_F = 4.4836$. Omitting x^3 the RSS increases from $RSS_F = 246.6$ to $RSS_R = 288.7$ so that the test statistic is

$$F_{obs} = \frac{288.7 - 246.6}{4.4836} = 9.3897 \qquad (\mathbf{1})$$

The critical value is the upper 5% value of $F_{1,55}$ which is 4.0162. Hence $F_{obs} > F_{crit}$ so H_0 is rejected and x^3 must be retained. **(1)**
Further testing is inappropriate and hence the best fitting polynomial model is the third order model $Y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \epsilon_i$. **(1)**

(iii) In any regression model the (unstandardised) residual for the i-the data point is

$$e_i = y_i - \hat{y}_i$$

where y_i is the observed response and $\hat{y}_i$ is the fitted value from the model. **(1)**
If the model is a good fit then the e_i should be approximately independently normally distributed with zero mean but different variances. **(1)**
(Equivalent description of standardised residuals, where variances are all 1, also sufficient for full marks).

(iv) (Explanation in terms of raw residuals e_i or standardised residuals r_i is sufficient for full marks).

The main diagnostic plot likely to be used here is that of residuals e_i versus fitted values $\hat{y}_i$. There should be a random scatter of residuals about zero throughout the range of fitted values. **(1)**

Alternatively, a plot of residuals e_i versus predictor x_i should also produce a random scatter of residuals about zero throughout the range of x, and should be quite similar because there is only one predictor in this case. **(1)**

A common problem detected by residual plots is when the variance of the residuals increases with fitted value, suggesting that the assumption of constant variance is incorrect. This is often solved by transformation of the response variable, e.g. to a log scale. (With standardised residuals variance should be constant, with unstandardised residuals variance will increase at extremes of x range, so increase with fitted value implies non-constant variance of ϵ_i in either case) **(1)**

In addition to the above, a normal probability plot and a histogram of the residuals can both be used to assess the assumption of normality. For a normal probability plot the ordered residuals are plotted against the cumulative distribution function of the standard normal distribution, so that a straight line plot supports normality. **(1)**

Residual plots are best suited to assessing whether the model fitted contains too few terms, such as fitting a quadratic (2nd order) where a cubic (third order) was needed, and whether the assumption of independent normally distributed errors is valid. **(1)**

If a quadratic model is fitted to a cubic relationship then there should be a pattern to the plot of residuals versus fitted values, where a model with one turning point is fitted to a relationship with two, so that the residuals should fall into four groups, mostly positive, then negative, then positive, then negative (or the reverse). **(1)**

Backward elimination cannot detect this.

In contrast, backward elimination can assess whether the model fitted contains too many terms, such as fitting a cubic where a quadratic was adequate, but cannot assess the necessity or otherwise of extra terms in the model beyond those which have been used and cannot assess the assumption of normal and constant variance errors, which are used in the calculations. **(1)**

If a cubic model is fitted to a quadratic relationship then the residual plot is unlikely to show very much, as the cubic term β_3 is likely to be small so the fit will look very much like a quadratic within the range of x values used. **(1)**

Backward elimination should be able to detect this, provided the assumption of normally distributed errors with constant variance is valid, by the test detecting that the extra parameter is close to zero and not statistically significant. **(1)**

(Marks are awarded whether the student makes the key points in the general or the specific case).

2. (i) The points can be made in any order, or mixed together as long as it is clear, the following are put under specific headings purely to clarify the main points that the students need to make.

   (a) Allocation of treatment to plot should be randomised, in the sense that the probability of a given plot receiving a given treatment or a given method should be equal for all treatments, methods and plots. **(1)**
   This increases the chances of a fair comparison between the treatments and methods, by reducing the chance of any systematic allocation method biasing the results due to any connections between the allocation method and some variable that might affect the performance of the crop in a plot. **(1)**
   In particular, in this case any systematic allocation method could easily end up allocating more of the 'better' plots to some treatments or methods than others. **(1)**

   (b) The obvious design here is a five pesticide treatments by five non-lethal methods factorial design, with one observation (plot) per treatment/method combination, which is a factorial design because each treatment/block combination appears. **(1)**
   If only one factor at a time were varied then each observation (plot) would compare only the treatments or only the methods, not both. In this factorial design each observation helps to compare both treatments and methods at once. **(1)**
   Hence the factorial structure allows both factors to be investigated at the same time using the same resources, hence obtaining greater precision of estimation with the same level of resources. **(1)**
   In this case the factorial design will therefore give more powerful tests to distinguish between different pesticide treatments and different non-lethal methods. **(1)**

   (c) A control group represents a baseline performance against which the other treatments are to be tested. **(1)**
   In this case the 'no pesticide' and 'standard pesticide' treatments are in a sense both control groups, and any concentration of pesticide would only be useful if its performance exceeded that of both control groups by a degree that is both statistically significant and practically relevant. **(1)**
   The researcher should be asked why there are two control groups as this is rather unusual, we would expect that the 'standard pesticide' treatment has already been shown to be better than the 'no pesticide' treatment, otherwise why is it the standard at all? **(1)**
   The researcher needs a good reason for the 'no pesticide' treatment, perhaps hinted at by the claim that the new pesticide is 'environmentally friendly', so it may be that performance inferior to the 'standard pesticide' may be acceptable if it is better than 'no pesticide'. If this is not the case then the researcher should be persuaded to abandon the 'no pesticide' treatment, leaving six plots for each of the remaining treatments rather than five (though of course the design then becomes less neat). **(1)**

   (ii) Blocking can be performed when experimental units (plots of land) can be divided into blocks where there is reason to believe that observations from within the same block will be more similar than those from different blocks. **(1)**

The question clearly implies that the plots of land differ and that information is available to assess and quantify this, so that data from past use of these plots for this (sort of) crop should be investigated for clear patterns that can be used to define the blocks. **(1)**

Ideally we hope that five blocks, each of five plots, will appear naturally, so that each treatment appears exactly once in each block, but there is no particular reason to expect this to happen, and the design would then depend on the exact details of the size and distribution of the block effect. However, provided the plots can be approximately ranked in terms of how well the crop is likely to do, the best five plots will form one block, the next best five another and so on. **(1)**

Randomisation to treatment will then be performed separately for each block, increasing the chance of treatments being fairly allocated to plots of land. **(1)**

Blocks can also be included as a term in the analysis, reducing the residual variance and hence increasing the power of all tests. **(1)**

(iii) A Latin square design is a way of testing a variable (treatment) with two blocking factors even though there are only enough experimental units for one blocking factor, e.g. a five-level treatment with two five-level blocks using 25 rather than 125 experimental units. Each treatment level occurs only once in each level of both blocking factors. **(1)**

For five treatments (A-E) and two blocking factors the allocation of treatments to the two blocks would look something like this:

Blocking	Blocking factor 2				
Factor 1	1	2	3	4	5
1	A	B	C	D	E
2	B	C	D	E	A
3	C	D	E	A	B
4	D	E	A	B	C
5	E	A	B	C	D

(1)

In the present case treatment and method are both 'treatments' and there is only one blocking factor, so in a strict sense this is not quite a Latin square, but we can take pesticide treatment as the treatment and method as one of the blocking factors and it works in much the same way, and the behaviour of the researcher suggests that pesticide treatment is more important than non-lethal method, so it is not unreasonable to view the latter as a block. **(1)**

In the example table above, the five blocks would be randomly allocated to the labels 1–5 for blocking factor 1, and the five non-lethal methods similarly randomly allocated to the labels 1–5 for blocking factor 2. **(1)**

3. (i) A possible model is

$$Y_{ij} = \mu + \tau_i + \beta_j + \epsilon_{ij} \qquad (\mathbf{1})$$

where Y_{ij} is the number of bankruptcies, μ is the overall mean, τ_i the effect of the i-th type of business, β_j the effect of the j-th region **(1)**
and $\epsilon_{ij} \sim N(0, \sigma^2)$ is the independent random variation term. **(1)**
The assumption of normal errors is clearly not entirely correct as the responses are counts rather than measurements. This also means that the assumption of constant variance is dubious because we would expect the variability to be related to the mean, such as in the Poisson where the mean equals the variance. However, the numbers are all fairly large and fairly similar, so that the approximation should be reasonable in this case. **(1)**

(ii)

$$\begin{aligned}
TSS &= 45659 - cfm = 45659 - \frac{945^2}{20} = 45659 - 44651.25 = 1007.75 \\
TypeSS &= \frac{274^2 + 241^2 + 204^2 + 226^2}{5} - cfm \\
&= 45169.80 - 44651.25 = 518.55 \quad (\mathbf{1}) \\
RegionSS &= \frac{196^2 + 166^2 + 194^2 + 197^2 + 192^2}{4} - cfm \\
&= 44820.25 - 44651.25 = 169.00 \quad (\mathbf{1}) \\
RSS &= TSS - TypeSS - RegionSS \\
&= 1007.75 - 518.55 - 169.00 = 320.20 \quad (\mathbf{1})
\end{aligned}$$

Hence

Source	Sum of squares	d.f.	Mean Square	F ratio
Types	518.55	3	172.85	6.48
Regions	169.00	4	42.25	1.58
Residual	320.20	12	26.68	
Total	1007.75	19		

(d.f.) **(1)**

(iii) To test $H_0 : \tau_1 = \tau_2 = \tau_3 = \tau_4$ versus $H_1 :$ not H_0 **(1)**
at the 5% level we compare $F_{obs} = 6.48$ to $F_{crit} = F_{3,12,0.95} = 3.490$ **(1)**.
Here $F_{obs} > F_{crit}$ so we reject H_0 at the 5% level and conclude that there is a difference in mean between the types of business. **(1)**

To test $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5$ versus $H_1 :$ not H_0 **(1)**
at the 5% level we compare $F_{obs} = 1.58$ to $F_{crit} = F_{4,12,0.95} = 3.26$ **(1)**.
Here $F_{obs} < F_{crit}$ so we do not reject H_0 at the 5% level and conclude that there is no evidence of a difference in mean between the regions. **(1)**

(iv) Credit is awarded for discussion of *statistical* issues such as:
The data set only shows that a greater absolute number of businesses of certain types are going bankrupt than others, but that there is no difference in absolute number between regions. **(1)**
However, this says nothing about the probability of different businesses going bankrupt, because the data do not give the number of businesses overall. To make meaningful conclusions we would need to know the number of businesses of each type in each region, and also which ones receive loans from this bank. **(1)**
However, even this would not necessarily give information on how easy it is to induce bankruptcy, as the data are purely observational and do not record different approaches used by the bank. **(1)**

(v) An interaction occurs when the effects of the levels of one factor differ depending on the level of another factor. In this case, this would mean that the effect of type of business differs between the regions, or equivalently that the effect of region differs between the business types. **(1)**
In order to estimate interactions we need more than one observation per cell (type/region combination) in order to be able to estimate residual variation separately from type and region effects, hence allowing the presence of interactions between factors to be distinguished from large residual variation. **(1)**
One obvious way of doing that in this case would be to have data for two or more years, with the year being a natural unit of measurement to smooth over seasonal effects, but the standard analysis would then require the assumption that the parameters have not changed over time. **(1)**

4. (a) A cusum chart plots the cumulative sum of the difference between actual and target values of some quantity against time or observation number. **(1)**

Hence for a target value of k for some quantity x, it plots $\sum_{i=1}^{t}(x_i - k)$ or $\sum_{i=1}^{t}(\bar{x}_i - k)$ versus $t = 1 \ldots$. **(1)**

In this case the target is $k = 100$ and the observations x_i are the observed lengths of items taken from the production line, so that we plot $\sum_{i=1}^{t}(x_i - 100)$ versus $t = 1 \ldots$. **(1)**

If the mean of the system really is $k = 100$ then the chart simply shows random variation about $k = 100$, so that $\sum_{i=1}^{t}(x_i - k)$ randomly varies about zero, **(1)**

but if it moves away from $k = 100$ then this should show up fairly quickly because every value collected contributes to the cumulative sum, driving the running mean away from $k = 100$ and hence the value of $\sum_{i=1}^{t}(x_i - k)$ away from zero. **(1)**

The steepness of the gradient of the graph indicates how far the mean is away from k while increasing steepness shows that the mean is moving further away from $k = 100$, but early values well away from k will stay in the calculation even after the system is brought under control, so the calculation may need to be reset sometimes. **(1)**

However, if the system is in control, but is producing items which are on average slightly longer than $k = 100$, then the cusum will show $\sum_{i=1}^{t}(x_i - k)$ steadily increasing against t even if the deviation from k is within acceptable tolerances (or similarly for the mean slightly below k). **(1)**

In comparison to non-cumulative plots of means, such as Shewhart charts, a cusum chart will tend to indicate a small change in mean more quickly, but may be slightly slower to pick up a sudden large change. **(1)**

(b) (i) We are given that the true proportion of faulty items in a batch is p, so if we assume that items are faulty independently of each other, each with probability p, then the number of faulty items in a batch will follow the binomial distribution. This is reasonable if items are sampled randomly from each batch. **(1)**

Scheme A: the number of faulty items is $X \sim Bi(30, p)$ and we accept the batch if $X \leq 2$. **(1)**

Putting $q = 1 - p$

$$\begin{aligned} P(\text{Accept}) = P(X \leq 2) &= q^{30} + 30pq^{29} + \frac{30 \times 29}{2}p^2q^{28} \\ &= q^{30} + 30pq^{29} + 435p^2q^{28} \quad \textbf{(1)} \end{aligned}$$

Scheme B: let the number of faulty items in stage j be $Y_j \sim Bi(20, p)$ for $j = 1, 2$ **(1)**

so that

$$\begin{aligned} P(\text{Accept}) &= P(Y_1 = 0) + P(Y_1 = 1) \times P(Y_2 \leq 2) + P(Y_1 = 2) \times P(Y_2 \leq 1) \\ &\quad + P(Y_1 = 3) \times P(Y_2 = 0) \end{aligned}$$

$$\begin{aligned} &= p_0 + p_1(p_0 + p_1 + p_2) + p_2(p_0 + p_1) + p_3 p_0 \\ &= p_0(1 + p_1 + p_2 + p_3) + p_1^2 + 2p_1 p_2 \quad (\mathbf{1}) \end{aligned}$$

where

$$\begin{aligned} p_0 = P(Y_j = 0) &= q^{20} \\ p_1 = P(Y_j = 1) &= 20pq^{19} \\ p_2 = P(Y_j = 2) &= \frac{20 \times 19}{2} p^2 q^{18} = 190 p^2 q^{18} \\ p_3 = P(Y_j = 3) &= \frac{20 \times 19 \times 18}{6} p^3 q^{17} = 1140 p^3 q^{17} \quad (\mathbf{1}) \end{aligned}$$

and hence

$$\begin{aligned} P(\text{Accept}) &= q^{20} + 20pq^{39} + 190p^2 q^{38} + 1140p^3 q^{37} + 400p^2 q^{38} + 7600p^3 q^{37} \\ &= q^{20} + 20pq^{39} + 590p^2 q^{38} + 8740p^3 q^{37} \quad (\mathbf{1}) \end{aligned}$$

(ii) Scheme A:

$$\begin{aligned} P(\text{Accept}) &= 0.95^{30} + 30 \times 0.05 \times 0.95^{29} + 435 \times 0.05^2 \times 0.95^{28} = 0.8122 \\ \Rightarrow P(\text{Reject}) &= 1 - 0.8122 = 0.1878 \quad (\mathbf{1}) \end{aligned}$$

Scheme B:

$$\begin{aligned} P(\text{Accept}) &= 0.95^{20} + 20 \times 0.05 \times 0.95^{39} + 590 \times 0.05^2 \times 0.95^{38} \\ &\quad + 8740 \times 0.05^3 \times 0.95^{37} \\ &= 0.3585 + 0.1353 + 0.2100 + 0.1638 = 0.8676 \\ \Rightarrow P(\text{Reject}) &= 1 - 0.8676 = 0.1324 \quad (\mathbf{1}) \end{aligned}$$

Hence B is less likely to reject a batch for this borderline figure of 5% faulty. However, it would require further algebra to check if this is true for values $p < 0.05$, so it cannot be stated that B is therefore better in terms of avoiding false positives, and any calculation of which method is better would also need to include the expected number of stages, and hence items sampled, for different values of p. **(1)**

(iii) The cost of sampling an item **(1)**
and the costs of false positive and false negative rejection of batches must both be defined. **(1)**