THE ROYAL STATISTICAL SOCIETY

2010 EXAMINATIONS – SOLUTIONS

HIGHER CERTIFICATE

MODULE 6

FURTHER APPLICATIONS OF STATISTICS

The Society provides these solutions to assist candidates preparing for the examinations in future years and for the information of any other persons using the examinations.

The solutions should NOT be seen as "model answers". Rather, they have been written out in considerable detail and are intended as learning aids.

Users of the solutions should always be aware that in many cases there are valid alternative methods. Also, in the many cases where discussion is called for, there may be other valid points that could be made.

While every care has been taken with the preparation of these solutions, the Society will not be responsible for any errors or omissions.

The Society will not enter into any correspondence in respect of these solutions.

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

Higher Certificate, Module 6, 2010. Question 1

(i) The "correction term" for the analysis of variance is $\frac{877.8^2}{25} = 30821.3136$.

The total sum of squares is therefore 32186.08 – 30821.3136 = 1364.7664. This has 25 – 1 = 24 df.

The sum of squares for methods is

$$\frac{213.6^2}{8} + \frac{185.0^2}{5} + \frac{170.8^2}{4} + \frac{170.8^2}{4} + \frac{137.6^2}{4} - 30821.3136 = 1046.5664,$$

with 4 df.

The residual sum of squares and degrees of freedom are obtained by subtraction.

Hence the completed analysis of variance table is as follows.

SOURCE OF VARIATION	DEGREES OF FREEDOM	SUM OF SQUARES	MEAN SQUARE	*F* value
Methods	4	1046.5664	261.6416	16.45
Residual	20	318.2000	15.91	$= \hat{\sigma}^2$
TOTAL	24	1364.7664		

The *F* value is referred to $F_{4,20}$ and is very highly significant: the upper 0.1% critical point is 7.10. Thus there is extremely strong evidence of differences among the method means.

The model that is the basis for this analysis is

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

where y_{ij} is the *j*th result using the *i*th method, μ is the overall population mean for the results, α_i is the population mean amount by which the results for the *i*th method differ from μ, and ε_{ij} is a random residual (error) term with ε_{ij} ~ ind N(0, σ^2) where σ^2 is a constant.

Solution continued on next page

(ii) The observed means for methods S and A are 26.7 and 37.0. These are the means of 8 and 5 observations respectively. The underlying variance of the difference in means is therefore $\{(1/8)+(1/5)\}\times\sigma^2$ where σ^2 is the variance underlying each observation. We estimate this by $\{(1/8)+(1/5)\}\times 15.91$ = 5.17075.

The double-tailed 5% point of t_{20} is 2.086.

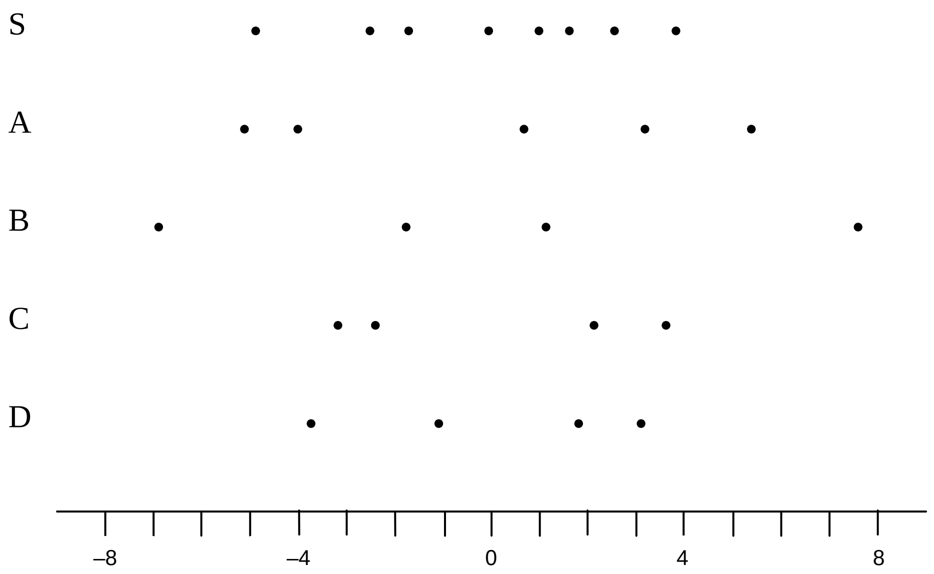
So a 95% confidence interval for the true mean difference (A – S) is

$$10.3 \pm 2.086\sqrt{5.17075} \;=\; (5.56,\; 15.04).$$

(iii) The residuals are as follows.

A:	–5.2, 0.7, –4.1, 3.2, 5.4
B:	–7.0, –1.8, 7.7, 1.1
C:	–2.5, –3.2, 3.6, 2.1
D:	–3.8, –1.1, 1.8, 3.1

(iv)



Constant variance seems unlikely (for example, the observations for B are much more spread out than those for C or D). Normality also seems unlikely – there is only a partial suggestion, for some of the methods, of clustering near 0 such as we would expect under Normality.

Higher Certificate, Module 6, 2010. Question 2

(i) The model is $Y = \beta X + \gamma Z$ and we have observations (x_i, y_i, z_i) for $i = 1, 2, \ldots, n$.

We minimise $S = \sum_{i=1}^{n} \left(y_i - \beta x_i - \gamma z_i \right)^2$.

We do this by setting derivatives equal to 0. (Strictly we should also check the second derivates, to ensure that we locate a minimum; this step is omitted here.)

$\frac{\delta S}{\delta \beta} = -2\Sigma x_i \left(y_i - \beta x_i - \gamma z_i \right)$, so we have $\Sigma x_i y_i = \hat{\beta} \Sigma x_i^2 + \hat{\gamma} \Sigma x_i z_i$.

$\frac{\delta S}{\delta \gamma} = -2\Sigma z_i \left(y_i - \beta x_i - \gamma z_i \right)$, so we have $\Sigma z_i y_i = \hat{\beta} \Sigma x_i z_i + \hat{\gamma} \Sigma z_i^2$.

These equations may be solved together as follows.

Multiply the first by Σz_i^2 and the second by $\Sigma x_i z_i$ to give

$$\Sigma x_i y_i \Sigma z_i^2 = \hat{\beta} \Sigma x_i^2 \Sigma z_i^2 + \hat{\gamma} \Sigma x_i z_i \Sigma z_i^2$$
$$\Sigma z_i y_i \Sigma x_i z_i = \hat{\beta} \left(\Sigma x_i z_i \right)^2 + \hat{\gamma} \Sigma z_i^2 \Sigma x_i z_i$$

and now subtract to obtain

$$\hat{\beta} = \frac{\Sigma x_i y_i \Sigma z_i^2 - \Sigma x_i z_i \Sigma z_i y_i}{\Sigma x_i^2 \Sigma z_i^2 - \left(\Sigma x_i z_i \right)^2}.$$

Similarly, multiply the first by $\Sigma x_i z_i$ and the second by Σx_i^2 and then subtract to obtain

$$\hat{\gamma} = \frac{\Sigma z_i y_i \Sigma x_i^2 - \Sigma x_i z_i \Sigma x_i y_i}{\Sigma x_i^2 \Sigma z_i^2 - \left(\Sigma x_i z_i \right)^2}.$$

(ii) (a) Simply replace Z by X^2 throughout. The estimators become

$$\hat{\beta} = \frac{\Sigma x_i y_i \Sigma x_i^4 - \Sigma x_i^3 \Sigma x_i^2 y_i}{\Sigma x_i^2 \Sigma x_i^4 - \left(\Sigma x_i^3 \right)^2}, \qquad \hat{\gamma} = \frac{\Sigma x_i^2 y_i \Sigma x_i^2 - \Sigma x_i^3 \Sigma x_i y_i}{\Sigma x_i^2 \Sigma x_i^4 - \left(\Sigma x_i^3 \right)^2}.$$

Solution continued on next page

(b) Set up columns of data (n rows for each) for Y, X and X^2, treating X^2 as a variable in its own right.

Use the computer program to fit the model $Y = \beta X + \gamma X^2$ (this will be achieved by <u>not</u> making use of whatever the "fit constant" option of the program is).

Also use the computer program to fit the model $Y = \alpha + \beta X + \gamma X^2$ (this time using the "fit constant" option, which may be the default). Perform the usual t test for $\alpha = 0$, using the information in the output (i.e. the value of the estimate of α and its standard error).

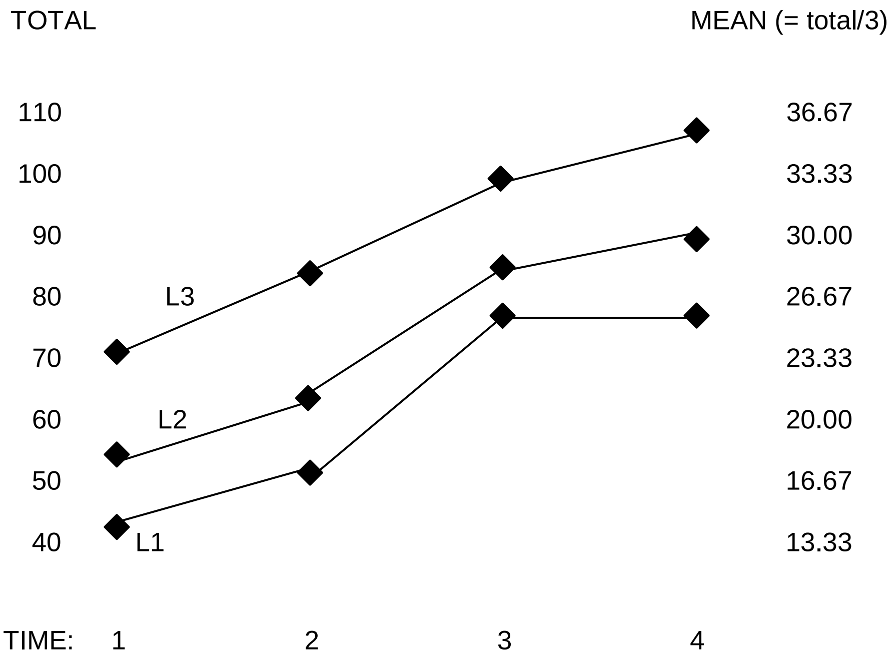
For each model, check the overall goodness of fit by the usual F test comparing the regression mean square with the residual mean square. The R^2 value may also be helpful.

Look at the residuals from each model. Compare the patterns (do they look like random Normal observations?) and the sizes of the residuals. Are there any signs of systematic patterns or of curvature?

Check any influential items of data that are identified in the output.

Higher Certificate, Module 6, 2010. Question 3

(i)



(ii) The "correction term" for the analysis of variance is $\frac{891^2}{36} = 22052.25$.

The total sum of squares is given in the question (1596.75) together with its number of degrees of freedom (35).

The sum of squares for blocks is

$$\frac{265^2}{12} + \frac{294^2}{12} + \frac{332^2}{12} - 22052.25 = 188.17 \quad \text{(with 2 df, given in the question).}$$

The sum of squares for time (T) is

$$\frac{166^2}{9} + \frac{196^2}{9} + \frac{258^2}{9} + \frac{271^2}{9} - 22052.25 = 834.08, \quad \text{with 3 df.}$$

The sum of squares for levels (L) is

$$\frac{245^2}{12} + \frac{289^2}{12} + \frac{357^2}{12} - 22052.25 = 530.67, \quad \text{with 2 df.}$$

Solution continued on next page

The total sum of squares for "treatments" is given in the question: 1375.42. So, by subtraction, the sum of squares for the T × L interaction is given by 1375.42 – 834.08 – 530.67 = 10.67, and this has 3 × 2 = 6 df.

The residual sum of squares and degrees of freedom are now obtained by subtraction.

Hence the completed analysis of variance table is as follows.

SOURCE OF VARIATION	DEGREES OF FREEDOM	SUM OF SQUARES	MEAN SQUARE	*F* VALUE
Blocks	2	188.17	94.085	62.4
T	3	834.08	278.027	184.5
L	2	530.67	265.335	176.1
T × L	6	10.67	1.778	1.2
Treatments	11	1375.42		
Residual	22	33.16	1.507	
TOTAL	35	1596.75		

(iii) The *F* values are referred to the respective *F* distributions. Those for blocks, T and L are extremely highly significant. That for the T × L interaction is not significant (the upper 5% critical point of $F_{6,24}$ is 2.51). We conclude that there is extremely strong evidence for overall differences between the blocks (so it was worthwhile to have carried out the experiment according to a randomised blocks design); also for differences among the means for time (T) and for levels (L), with no evidence for any interaction between these effects.

Inspection of the data, and of the graph drawn in part (i), suggests that there is an increase in the yield from L1 to L2 to L3, at all times; and an increase in the yield from T1 to T2 to T3 to T4, for all levels (in this data set, level L1 actually remains the same from T3 to T4; but there is no evidence for any real interaction).

Higher Certificate, Module 6, 2010. Question 4

Let X be the number of faulty items in a sample.

(i) In Scheme I, $X \sim \mathrm{B}(24, 0.1)$. $P(\text{reject}) = P(X \geq 3) = 1 - P(X = 0, 1 \text{ or } 2)$

$$= 1 - \{(0.9)^{24} + 24(0.1)(0.9)^{23} + (24 \times 23/2)(0.1)^2(0.9)^{22}\} = 0.4357.$$

(ii) In Scheme II, we work with $X \sim \mathrm{B}(12, 0.1)$.

(a) P(accept batch after taking only the first sample)

$$= P(X = 0) = (0.9)^{12} = 0.2824.$$

(b) P(reject batch after taking only the first sample)

$$= P(X \geq 3) = 1 - P(X = 0, 1 \text{ or } 2)$$

$$= 1 - \{(0.9)^{12} + 12(0.1)(0.9)^{11} + (12 \times 11/2)(0.1)^2(0.9)^{10}\}$$

$$= 0.1109.$$

(c) The batch is rejected after taking both samples when

either there is 1 faulty in the 1st sample and ≥ 2 faulty in the 2nd

or there are 2 faulty in the 1st sample and ≥ 1 faulty in the 2nd.

So the required probability is

$$P(X = 1)P(X \geq 2) + P(X = 2)P(X \geq 1)$$

$$= \{12(0.1)(0.9)^{11}\}\{1 - [(0.9)^{12} + 12(0.1)(0.9)^{11}]\} + \{(12 \times 11/2)(0.1)^2(0.9)^{10}\}\{1 - (0.9)^{12}\}$$

$$= 0.1284 + 0.1651 = 0.2935.$$

(iii) The overall probability of rejecting the batch using Scheme II is therefore 0.1109 + 0.2935 = 0.4044. Scheme I gave the slightly higher probability 0.4357. See comments at the end of part (iv).

Solution continued on next page

(iv) The results in parts (a) and (b) of (ii) show that the probability of making a decision after only one sample under Scheme II is 0.2824 + 0.1109 = 0.3933. So the probability of needing the second sample is 1 – 0.3933 = 0.6067.

$\therefore$ average sample size under Scheme II

$$= (12 \times 0.3933) + (24 \times 0.6067) = 19.28.$$

Thus Scheme II gives an average sample size approaching 5 fewer than the fixed sample size of Scheme I. Further (see part (iii)), the probability of rejecting a batch under Scheme II is not much less than that under Scheme I. So Scheme II would be likely to be preferred – on average the sample size is smaller, and there is only a quite small difference in the rejection probability.