THE ROYAL STATISTICAL SOCIETY

2008 EXAMINATIONS – SOLUTIONS

HIGHER CERTIFICATE

(MODULAR FORMAT)

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, 2008. Question 1

(i) The total sum of squares for the analysis of variance table is given in the question: 1583.1384. This has 25 – 1 = 24 df.

The sums of squares for rows and columns are also given in the question. Each has 5 – 1 = 4 df.

The grand total is 738.12. The "correction factor" is $\frac{738.12^2}{25} = 21792.84538$.

So the SS for treatments, also with 5 – 1 = 4 df, is

$$\frac{107.02^2}{5} + ... + \frac{192.67^2}{5} - 21792.84538 \;=\; 1317.5627 .$$

The residual SS and df 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
Rows	4	45.3682	11.342	0.67
Columns	4	17.9588	4.490	0.27
Treatments	4	1317.5627	329.391	19.54
Residual	12	202.2487	16.854	$= \hat{\sigma}^2$
TOTAL	24	1583.1384		

[Note. The *F* values for rows and columns are not significant; that for treatments is very highly significant.]

(ii) The observed means for groups A and B are 21.40 and 36.22. Each of these is the mean of 5 observations. The underlying variance of the difference in means is $(2/5)\times\sigma^2$ where σ^2 is the variance underlying each observation. We estimate this by $(2/5)\times 16.854 = 6.7416$.

The double-tailed 5% point of t_{12} is 2.179.

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

$$14.82 \pm 2.179\sqrt{6.7416} \;=\; (9.16,\; 20.48).$$

As the interval does not contain 0, we may conclude that the mean for B is significantly different from that for A. Storage time B appears to lead on average to greater weight loss than storage time A. We are "95% confident" (in the usual interpretation of confidence intervals) that this greater weight loss is between about 9.2 and 20.5 per cent.

Solution continued on next page

(iii) We compare the mean for C (21.11; five observations) with the overall mean for D and E ((30.36 + 38.53)/2 = 34.45; ten observations). The underlying variance of the difference between these means is $[(1/5) + (1/10)]\times\sigma^2$ where σ^2 is the variance underlying each observation. We estimate this by $(3/10)\times 16.854 = 5.0562$.

So the test statistic is $\dfrac{13.34}{\sqrt{5.0562}} = 5.93$, which we refer to t_{12}.

This is extremely highly significant (the double-tailed 0.1% point of t_{12} is 4.318). There is overwhelming evidence against the null hypothesis here; it appears that there is an effect of storage time.

(iv) We compare the mean for D (30.36; five observations) with that for E (38.53; five observations). The underlying variance of the difference between these means is $(2/5)\times\sigma^2$ where σ^2 is the variance underlying each observation. We estimate this by $(2/5)\times 16.854 = 6.7416$.

So the test statistic is $\dfrac{8.17}{\sqrt{6.7416}} = 3.15$, which we refer to t_{12}.

This is highly significant (the double-tailed 1% point of t_{12} is 3.055). There is strong evidence against the null hypothesis here; it appears that there is an effect of covering.

[Note. In parts (iii) and (iv), there is also an argument that the interpretation should be one-sided (i.e. an alternative hypothesis of "weight loss is less"). The conclusions would be that storage for a shorter storage time (i.e. C) appears to be better and, of the two longer times, D (use of protective covering) appears to be better than E (no protective covering).]

Higher Certificate, Module 6, 2008. Question 2

(a) The model is $Y = \alpha + \beta_1 X_1 + \beta_2 X_2$ and we have observations (y_i, x_{1i}, x_{2i}) for $i = 1, 2, \ldots, n$.

We minimise $S = \sum_{i=1}^{n} \left(y_i - \alpha - \beta_1 x_{1i} - \beta_2 x_{2i} \right)^2$.

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.)

$\frac{\delta S}{\delta \alpha} = -2\Sigma \left(y_i - \alpha - \beta_1 x_{1i} - \beta_2 x_{2i} \right)$, so we have $0 = \Sigma y_i - n\hat{\alpha} - \hat{\beta}_1 \Sigma x_{1i} - \hat{\beta}_2 \Sigma x_{2i}$.

$\frac{\delta S}{\delta \beta_1} = -2\Sigma x_{1i} \left(y_i - \alpha - \beta_1 x_{1i} - \beta_2 x_{2i} \right)$, so $0 = \Sigma x_{1i} y_i - \hat{\alpha} \Sigma x_{1i} - \hat{\beta}_1 \Sigma x_{1i}^{\,2} - \hat{\beta}_2 \Sigma x_{1i} x_{2i}$.

$\frac{\delta S}{\delta \beta_2} = -2\Sigma x_{2i} \left(y_i - \alpha - \beta_1 x_{1i} - \beta_2 x_{2i} \right)$, so $0 = \Sigma x_{2i} y_i - \hat{\alpha} \Sigma x_{2i} - \hat{\beta}_1 \Sigma x_{1i} x_{2i} - \hat{\beta}_2 \Sigma x_{2i}^{\,2}$.

Thus the normal equations are $\begin{cases} \Sigma y_i = n\hat{\alpha} - \hat{\beta}_1 \Sigma x_{1i} - \hat{\beta}_2 \Sigma x_{2i} \\ \Sigma x_{1i} y_i = \hat{\alpha} \Sigma x_{1i} - \hat{\beta}_1 \Sigma x_{1i}^{\,2} - \hat{\beta}_2 \Sigma x_{1i} x_{2i} \\ \Sigma x_{2i} y_i = \hat{\alpha} \Sigma x_{2i} - \hat{\beta}_1 \Sigma x_{1i} x_{2i} - \hat{\beta}_2 \Sigma x_{2i}^{\,2} \end{cases}$

Solution continued on next page

(b) To use backwards elimination, it is convenient to set up a table similar to that shown below. In each row of the table, the residual SS is calculated as the total SS (4.9150, given in the question) minus the regression SS (also given in the question). Note that there are 8 observations, and thus 7 df for the total SS.

Variables included	Regression SS	Residual SS and df	
x_1 and x_2	4.1139	0.8011	5
x_1	0.1486	4.7664	6
x_2	4.0439	0.8711	6

Note: total SS = 4.9150

The residual mean square from the full model, i.e. the model with x_1 and x_2, is 0.8011/5 = 0.1602 with 5 df. We use this initially.

The smallest change from the full model is clearly that which keeps x_2 (i.e. omits x_1), reducing the regression SS by 4.1139 – 4.0439 = 0.07 [or, equivalently, increasing the residual SS by 0.07]. To check whether this is a significant change, we compare it with the current residual, referring the result to $F_{1,5}$. We get

$$\frac{0.07}{0.1602} = 0.44,$$

which is clearly not significant on $F_{1,5}$. This means that the model sum of squares has not been reduced significantly, so we may omit x_1 without any serious change in the fit of the model.

We now consider this new model (i.e. containing x_2 but not x_1) and consider whether x_2 can also be omitted, leaving a model with (possibly a constant term and) random error only. Here we consider

$$\frac{4.0439}{0.8711/6} = 27.85,$$

and refer this to $F_{1,6}$. This is extremely highly significant (the upper 0.5% point is 18.63). This means that x_1 should be included in the model.

It appears that this cholesterol level depends on age but not on weight.

Higher Certificate, Module 6, 2008. Question 3

Part (a)

(i) If similar data sets are obtained from two different sources (for example, the same experiment carried out at two different sites), or there is a qualitative factor involved (for example a difference in the level of response between males and females given the same treatment), then a (dummy) indicator variable can be useful.

For illustration, suppose there is just one predictor variable x_1. A likely regression model is one having two parallel lines a distance (vertical) apart d. Introduce a binary variable x_2 (i.e. a variable that takes just two values, say 0 and 1). For example in the male/female situation, x_1 might be taken equal to 1 for males and 0 for females. The model becomes $y = \alpha + \beta_1 x_1 + \beta_2 x_2$, which we fit as a multiple regression. The value of β_2 is the distance d because x_2 goes from 0 to 1 as the units change from the one group (say the females) to the other (males), and y goes up by d. β_1 is the common slope of the two lines. α is the intercept when $x_1 = x_2 = 0$.

(ii) In a quadratic regression with (general) model $y = a + bx + cx^2$ [+ error], it may be thought that (the expected value of) y must be zero when x is zero, in which case a in the model is 0. An example in plant breeding might be where y is the height to which a plant grows and x represents the level of a fertiliser without which the plant will not grow at all. Another example might be in a food process that requires cooking where y represents some characteristic of the prepared food and x the temperature to which the food is raised in cooking, the zero level being taken as room temperature or some other suitable minimum.

A maximum value of y in such a model corresponds with a turning point on the response curve $y = bx + cx^2$. This can be located by ordinary calculus: $dy/dx = b + 2cx$, which equals zero when $x = -b/(2c)$. This is estimated by inserting the estimated values of b and c, and the estimated value of y is then found simply by inserting this value of x into the estimated regression. Strictly speaking it should be confirmed that it is indeed a maximum that has been located.

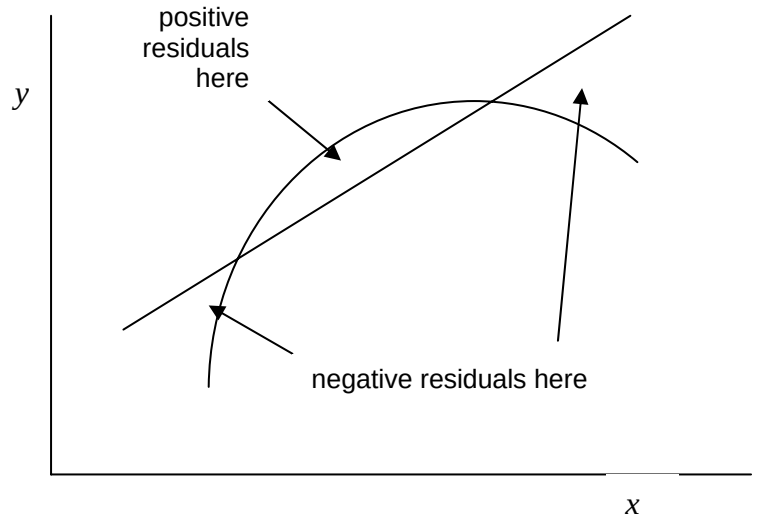
If the estimated value of x obtained as above is not within the range of the available data, we are faced with the danger of extrapolation: we do not know whether the regression relation still holds at this value of x. This should be made clear in any report.

Solution continued on next page

Part (b)

There should be no systematic pattern of variation in the residuals. Various plots and summaries are commonly available in computer package output.

(1) If the correct model is quadratic, whereas only simple linear regression has been fitted, the residuals will not change sign purely at random as x increases.



(2) If one of the scales should have been logarithmic (so that the model should, for example, have been of the form $y = \log x$), a fan shape of the residuals will be seen when they are plotted against either the fitted or the observed values.

(3) Skewness of residuals may also show up in such a plot. The usual simple assumptions then do not hold; a transformation may be required.

(4) Lack of Normality of the residuals can be detected by a Normal probability plot.

(5) The situation in part (a)(i) might be discovered in a plot of the residuals by groups.

(6) The largest residuals indicate possible outliers. These are typically flagged in computer output. They should be explored for any systematic patterns or one-off explanations.

(7) "Influential" points are those that can substantially alter the slope of a line (or perhaps give a strong indication that the model is inadequate). These are also often flagged in computer output. They should be investigated thoroughly.

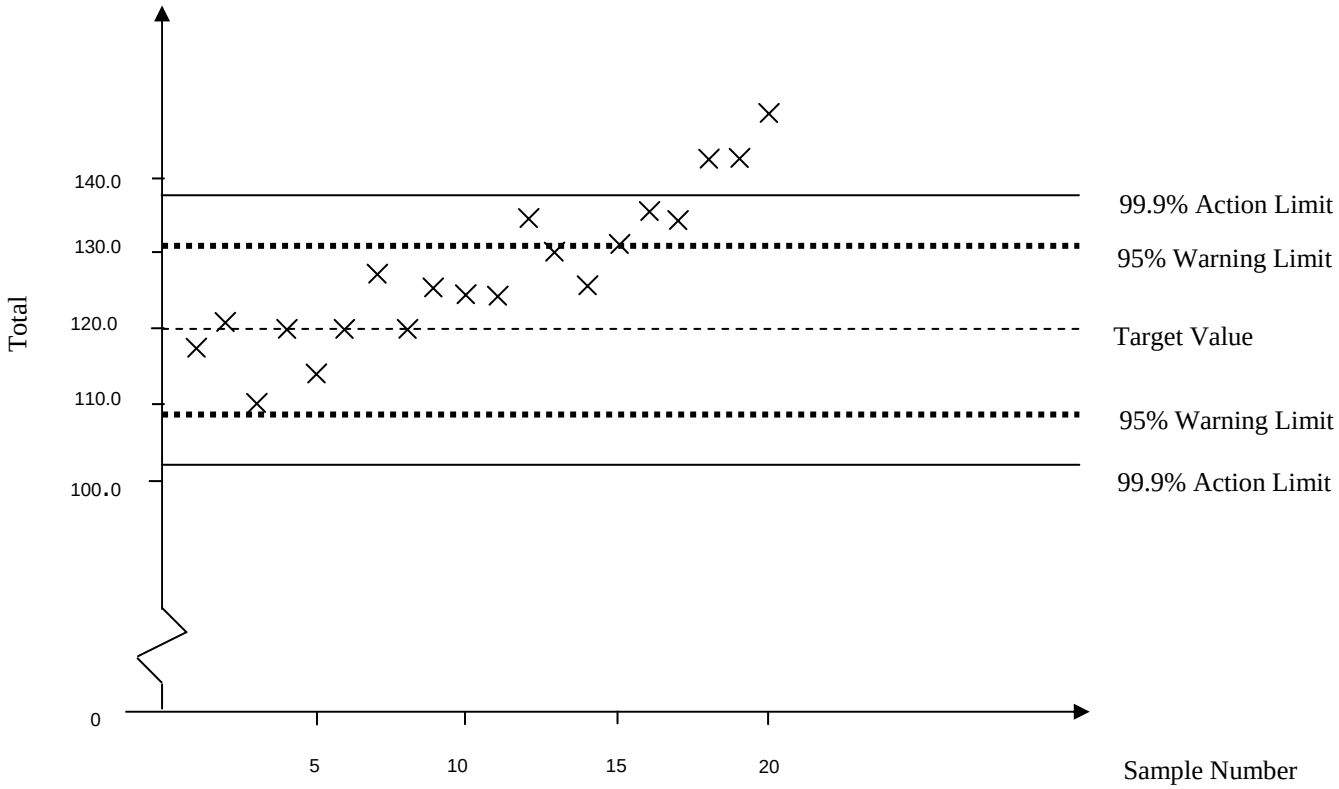
Higher Certificate, Module 6, 2008. Question 4

Warning limits for totals of samples of size 3 are "target ± 1.96($\sigma\sqrt{3}$)" and action limits at "target ± 3.29($\sigma\sqrt{3}$)" using the conventional UK Standard values, with probability levels 95% and 99.9%.

In the current US system and in "six sigma" work, factors of 2 and 3 are used instead of 1.96 and 3.29. These are also used in some computer packages. They are acceptable in candidates' solutions. They lead to similar inferences.

The target is 40 for an individual resistor (i.e. 120 for the total of three), and $\sigma = 3.2$. Hence the warning limits are at 109.14 and 130.86, and the action limits are at 101.76 and 138.24. The totals are plotted on the chart as shown below. Alternatively and equivalently, the sample means could be plotted with warning limits 36.38 and 43.62 and with action limits at 33.92 and 46.08.

Control Chart for totals of three values

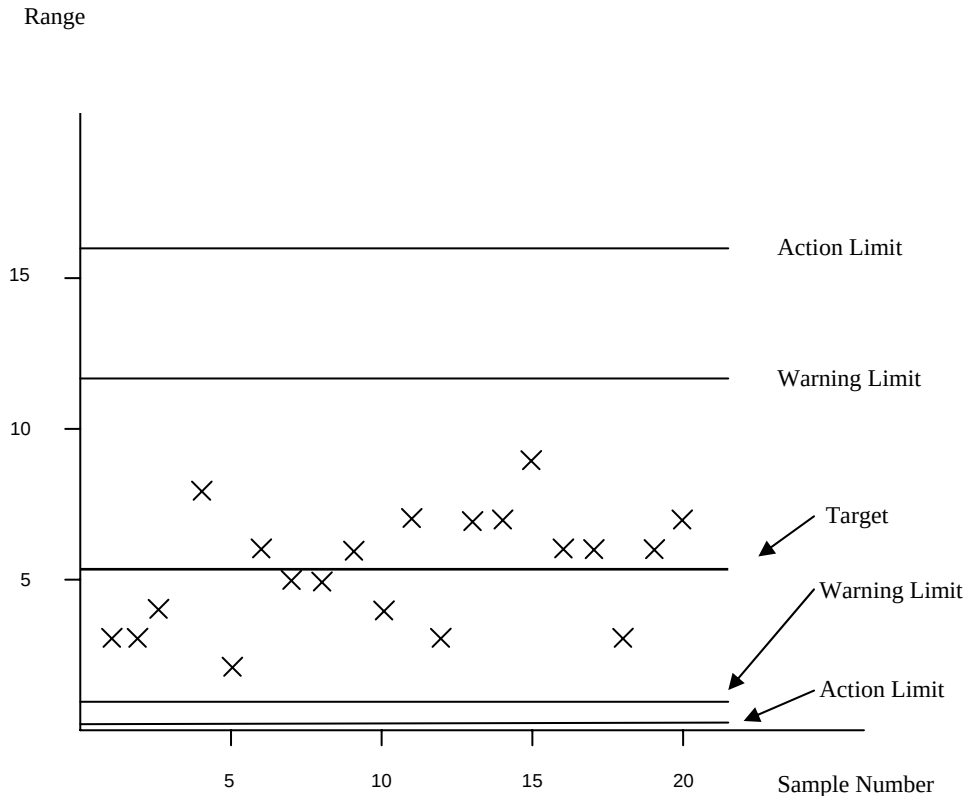


No totals fall below the lower warning limit. The total for the 12th batch is above the upper warning limit, but the totals for batches 13 and 14 are not, so the process is not yet stopped. However, the totals for all batches from 15 onwards are above the warning limit and those for 18, 19 and 20 are above the action limit. So we might reasonably conclude that the process mean was out of control at any time from batch 16 onwards, and certainly conclude this no later than batch 18.

Solution continued on next page

To construct a chart for the range, we first calculate the average range which is 107/20 = 5.35. Using the factors given in the question, the warning limits are therefore at 0.963 and 11.61 and the action limits at 0.214 and 16.00. [Note. The range can also be found from the standard deviation, which is given in the question, but the factors for doing that are not given.] The ranges are plotted as shown below.

Control Chart for ranges of three values



None of the ranges fall outside the warning limits, so there seems no question of the variability of the process being out of control.