THE ROYAL STATISTICAL SOCIETY

2009 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, 2009. Question 1

(i) Blocking removes a (known or suspected) source of systematic variation among the experimental units that are available. This allows a set of experimental results to be obtained that will be much more useful in investigating the effects which are of actual interest. A more precise comparison of treatment means is possible because the residual error estimate will reflect only random variation and not be inflated by the systematic variation.

(a) Blocking would be likely to be required in a field experiment with soil or climatic variation; or in a medical trial carried out at a number of different hospitals; or in an industrial experiment spread over several days.

(b) Blocking would be unlikely to be necessary in an experiment in closed, carefully controlled environmental conditions; in a small experiment at a single centre carried out by the same doctors/nurses over a short period on a homogeneous group of patients; or in a pilot trial on a small scale in industry.

(ii) [The solution covers parts (a) and (b) together.]

The "correction factor" for the analysis of variance is $\frac{265^2}{20} = 3511.25$.

The total sum of squares for the table is therefore 3611 – 3511.25 = 99.75. This has 20 – 1 = 19 df.

The sum of squares for blocks is $\frac{56^2}{4} + ... + \frac{60^2}{4} - 3511.25 = 30.00$, with 4 df.

For the main effect of A, note first that the totals for the low and high levels of A are 133 and 132 respectively, each being the total of 10 observations. So the sum of squares for A is

$$\frac{133^2}{10} + \frac{132^2}{10} - 3511.25 = 0.05, \text{ with 1 df.}$$

Similarly, the sum of squares for B is $\frac{130^2}{10} + \frac{135^2}{10} - 3511.25 = 1.25$, with 1 df.

Solution continued on next page

The sum of squares for the AB interaction can be obtained directly, or can be found by first obtaining the overall sum of squares for treatments and then subtracting the sums of squares for A and for B.

The calculation to obtain it directly first requires the totals when A and B are either both at the low level or both at the high level (this total is 58 + 60 = 118) and when one of them is at the high level and the other at the low level (this total is 72 + 75 = 147). The sum of squares for AB is then found as

$$\frac{118^2}{10}+\frac{147^2}{10}-3511.25=42.05.$$

The alternative method first requires the overall treatments sum of squares which is $\frac{58^2}{5}+...+\frac{72^2}{5}-3511.25=43.35$. Subtracting the sums of squares for A and for B gives 42.05, as before.

This sum of squares has 1 (= 1 × 1) df.

The residual sum of squares and its number of degrees of freedom can now be found 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	4	30.00	7.50	3.41
A	1	0.05	0.05	0.02
B	1	1.25	1.25	0.57
A × B	1	42.05	42.05	19.11
Residual	12	26.40	2.20	$= \hat{\sigma}^2$
TOTAL	19	99.75		

The upper 5% critical point of $F_{4,12}$ is 3.26. The upper 0.1% critical point of $F_{1,12}$ is 18.64. Thus there is evidence of a difference between blocks and extremely strong evidence of an AB interaction.

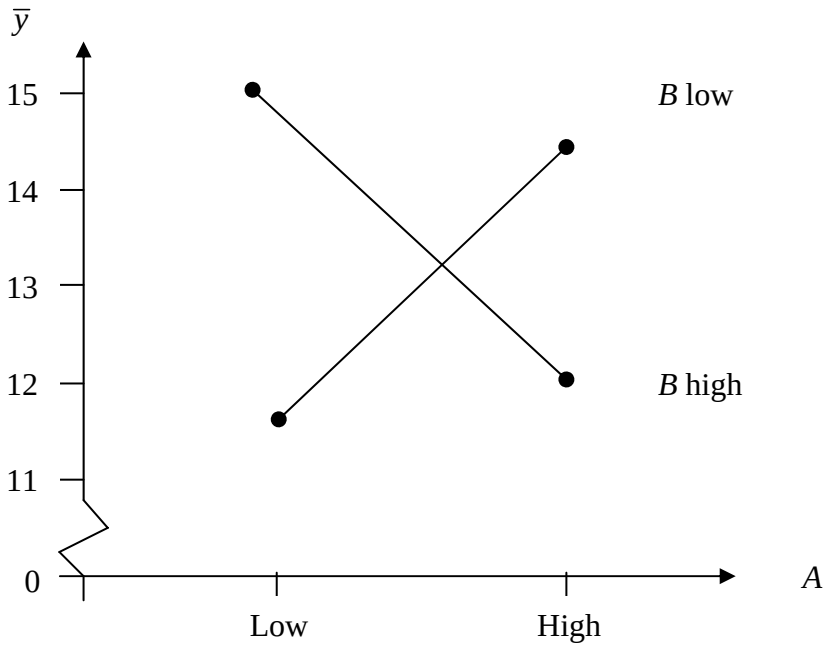
There is no evidence for separate main effects of A or B, but these are of no real value in interpreting the results because of the interaction.

Solution continued on next page

To study the interaction, we draw up a table of the two-way means:

	B low	B high
A low	11.6	15.0
A high	14.4	12.0

These may conveniently be plotted on a diagram:



The two factors seem to be working against each other. As high values of *y* are desirable, we should *either* use the high level of A with the low level of B *or* use the low level of A with the high level of B. It would not be sensible to use *both* A and B at the high level, or *both* of them at the low level.

It is not obvious what further work might be done. Perhaps the effects of using higher levels of one of the factors could be explored, or lower levels of one of the factors. In either case, the interactive effect must be kept in mind when analysing and interpreting the results. Another possibility might be to explore whether there is a best combination with both A and B at intermediate levels between the present "high" and "low" levels.

Higher Certificate, Module 6, 2009. Question 2

(i) At each step in the decision process, the probability of accepting the batch is $P(0)$ and the probability of taking a further sample is $P(1)$. Hence the batch is accepted after k steps if there was one faulty item at each step from 1 to $(k-1)$ inclusive and none faulty at step k.

So $P(\text{accept at step } k) = P(0).\{P(1)\}^{k-1}$.

The total probability of acceptance is thus $\sum_{k=1}^{\infty} P(0)\{P(1)\}^{k-1}$. This sum is simply $\dfrac{P(0)}{1-P(1)}$ (noting that $P(1) < 1$).

Similarly, the probability of rejection of a batch at step k is $P(2).\{P(1)\}^{k-1}$ where $P(2)$ denotes the probability of 2 *or more* defectives in a sample, and the total probability of rejection is

$$\sum_{k=1}^{\infty} P(2)\{P(1)\}^{k-1} = \frac{P(2)}{1-P(1)} = \frac{1-P(0)-P(1)}{1-P(1)}.$$

For convenience, denote these probabilities by p_a and p_r respectively, noting that $p_a + p_r = 1$. If now N is the random variable counting the number of batches inspected up to (and including) the first that is rejected, we have $P(N = 1) = p_r$, $P(N = 2) = p_a p_r$, $P(N = 3) = p_a^2 p_r$ and so on, i.e. N has a geometric distribution.

We require $E(N)$ which is (either by quoting the mean of a geometric distribution or by explicit evaluation) $1/p_r = \{1 - P(1)\}/\{1 - P(0) - P(1)\}$.

(ii) For the given case $n = 100$ and $p = 0.02$, we have $P(0) = (0.98)^{100} = 0.1326$ and $P(1) = 100(0.02)(0.98)^{99} = 0.2706$.

Hence $P(0)/\{1-P(1)\} = \dfrac{0.1326}{0.7294} = 0.1818$

and

$$\frac{1-P(1)}{1-P(0)-P(1)} = \frac{0.7294}{1-0.4032} = 1.22.$$

We have a large sample size and a very restrictive rule for acceptance; hence the probability of acceptance is low. Rejection occurs quite frequently, so the "average run length" is small.

Higher Certificate, Module 6, 2009. Question 3

(a) A Shewhart $\bar{x}$-chart is based on plotting successive means, of samples of n items, and action is based on individual means going outside action limits. The various rules that have been suggested for stopping a process when several successive means go outside warning limits do bring several means into the process, but Cusum charts extend this idea.

Suppose that the target value of x is m, and values of x are available at times 1, 2, 3, … . For each time i, the difference $x_i - m$ is calculated. Then $\sum_{i=1}^{r}(x_i - m)$ is plotted in the vertical direction against r.

This gives a plot of the "cumulative sum" of these differences. If the mean of x is equal to the target value m, then the plot should fluctuate around 0 in a random manner. But if it shows a steady underlying increase (or decrease), then the process is departing from being in control. The gradient of the (linear component of the) underlying trend shows how rapidly the process is going out of control, but the pattern is easy to see anyway.

Note that if the process is initially not giving average value m, then the initial average should be used in the Cusum calculation to avoid an initial slope up or down; but a Cusum chart may not be worth drawing until the required target has been obtained.

Cusums will show changes in mean almost immediately, an advantage over $\bar{x}$-charts, and will also pick up small changes from m more quickly. But large deviations show up immediately on an $\bar{x}$-chart, whereas a Cusum needs two or three observations to confirm a change. There are processes such as V-masks which can be used to assist in deciding when a change really has occurred.

Solution continued on next page

(b) (i) Treatments must not be allocated systematically to experimental units, nor by personal choice, because this leaves open the possibility of bias through deliberately avoiding allocating a treatment to the worst units (eg in a medical trial) or to the best. Bias can also arise accidentally: for example, in a field experiment, any systematic arrangement of treatments could be in phase with soil or climatic trends. Randomisation avoids this if strictly applied. Experiments without proper randomisation can be criticised (whether justifiably or not) if the results are "unexpected" by the critic. There is also the point that randomisation helps to validate results without the assumption of Normality for residual (error) terms.

(ii) Blinding is when people conducting a trial do not know which of the experimental treatments is being given to which experimental unit. This applies both to people actually administering treatments, eg a nurse in a clinical trial, and to people in charge of the experiment, eg the doctor supervising it. Blinding removes the danger of deliberate choice (see (b)(i)), provided that all treatments are indistinguishable from one another (eg all pills look exactly alike). The allocation to the units and, where necessary, the preparation of the treatments, are done by people not directly involved in running the experiment. Any labelling of treatments and/or units is done in some anonymous way.

Higher Certificate, Module 6, 2009. Question 4

Part (a)

(i) Examining residuals is the main way of assessing goodness of fit, and a dot plot of all the residuals will indicate whether there are (or may be) any outliers and whether the residuals could reasonably be a sample from a Normal distribution. Symmetry and central tendency might, for instance, be visually checked for.

(ii) A plot of residuals against fitted values is very useful. The fitted values are not correlated with the residuals, but the observed values are. There should be random scatter over the graph, without any pattern and without any evidence that variability changes with the size of the fitted value. Some of the same comments as in (i) also apply.

(iii) A Normal probability plot of the residuals can give a visual check on whether the assumption of Normality appears reasonable and can identify "unusual" residuals that might be checked to see whether they represent outliers.

(iv) Plots of residuals against predictor variables can show whether the right terms have been included – for example, to indicate whether x^2 should have been present in the model as well as x. Interpretation of such a plot depends on looking for any pattern in it.

(v) This plot checks for serial correlation, i.e. whether the assumption of (successive) independence might not be true.

Solution continued on next page

Part (b)

There are 13 observations, so there are 12 df for the total SS. The full model contains 4 variables, so the residual for the full model has 8 df.

Thus the analysis of variance for the full model is

	df	Sum of Squares	Mean Square
Regression	4	2667.90	
Residual	8	47.86	5.98
TOTAL	12	2715.76	

To use backwards elimination, we first make the smallest change from the full model by omitting x_3. The residual SS with this omitted is increased from 47.86 to 47.97, whereas omitting any other single variable would give a larger residual SS. To check whether this is a significant change, we compare it with the current residual, referring the result to $F_{1,8}$. We get

$$\frac{47.97 - 47.86}{5.98} = 0.02,$$

which is clearly not significant on $F_{1,8}$. This means that the model sum of squares has not been reduced significantly, so we may omit x_3 without any serious change in the fit of the model.

We now take this new model (i.e. containing x_1, x_2 and x_4) and consider whether one of these variables can also be omitted. The smallest increase in residual is by omitting x_4. To check whether this is a significant change, we again compare with the current residual, now referring the result to $F_{1,9}$. We get

$$\frac{57.90 - 47.97}{47.97/9} = 1.86,$$

and this is clearly not significant on $F_{1,9}$. This means that the model sum of squares has again not been reduced significantly, so we may omit x_4 as well as x_3 without any serious change in the fit of the model.

Proceeding similarly, we take the model containing x_1 and x_2 only and consider whether either of these can be omitted. The smallest increase in residual is by omitting x_1. To check whether this is a significant change, we again compare with the current residual, now referring the result to $F_{1,10}$. We get

$$\frac{906.34 - 57.90}{57.90/10} = 146.54,$$

and refer this to $F_{1,10}$. This is overwhelmingly significant and we very strongly conclude that x_1 should be included in the model (and therefore x_2 should be too, as this would lead to an even more highly significant result).

So the resulting model is $Y = a + b_1x_1 + b_2x_2$.