THE ROYAL STATISTICAL SOCIETY

HIGHER CERTIFICATE EXAMINATION

NEW MODULAR SCHEME

introduced from the examinations in 2007

MODULE 6

SPECIMEN PAPER A

AND SOLUTIONS

The time for the examination is 1½ hours. The paper contains four questions, of which candidates are to attempt **three**. Each question carries 20 marks. An indicative mark scheme is shown within the questions, by giving an outline of the marks available for each part-question. The pass mark for the paper as a whole is 50%.

The solutions should not be seen as "model answers". Rather, they have been written out in considerable detail and are intended as learning aids. For this reason, they do not carry mark schemes. Please note that in many cases there are valid alternative methods and that, in 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 the questions and solutions, the Society will not be responsible for any errors or omissions.

The Society will not enter into any correspondence in respect of the questions or solutions.

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

© RSS 2006

1. An experiment was performed to investigate the clotting time (in minutes) of plasma from 8 subjects treated by three different methods. The resulting data are presented in the table below.

Subject	*Method 1*	*Method 2*	*Method 3*
1	6.8	8.3	8.1
2	9.7	10.0	11.1
3	8.3	8.5	10.0
4	9.0	7.9	9.6
5	11.0	10.8	11.1
6	12.4	12.6	14.5
7	8.8	8.4	10.0
8	12.6	12.8	12.5

(i) Treating the subjects as blocks, analyse the data using analysis of variance. State the assumptions necessary for this analysis to be valid and provide a brief interpretation of your findings.

[Note. The sum of all the observations is 244.8 and the sum of their squares is 2585.22.]

(15)

(ii) Suppose that both the "method" sum of squares and the total sum of squares you calculated in part (i) had resulted from a similar experiment, which used 24 different subjects, eight being randomised to each method, rather than blocking on 8 subjects. Re-analyse the data on this basis and compare your conclusion with that from part (i). What impact has blocking had on the power to detect a difference between methods?

(5)

2. (i) From a very large batch of mass produced widgets, a simple random sample of 20 is selected and each widget in the sample is inspected to see whether or not it meets the quality requirement. The batch is accepted if the sample contains 0 or 1 defective widgets.

In a simple random sample, every widget has the same probability of being selected for the sample. Calculate the exact probability of accepting a batch if it contains a proportion p of defectives, for p = 0.01, 0.05, 0.1.

(4)

(ii) The above scheme is now modified in that, if the sample contains more than 2 defectives the batch is still rejected, but if the sample of 20 contains just 2 defectives, a further random sample of size 20 is selected, which may be assumed independent of the first sample. If the second sample contains no defectives then the batch is accepted, otherwise the batch is rejected.

List the possible numbers of defectives in the two samples which lead to acceptance of the batch, and hence calculate the probability of accepting a batch, for p = 0.01, 0.05, 0.1.

(7)

(iii) Rejected batches are subject to 100% inspection and all defectives are removed. If the batch size is 1000, calculate the expected total number of items inspected in the two sampling schemes in (i) and (ii), for each of the values of p given. Comment briefly on your results.

(9)

3. (i) A sample of companies from a business register has been taken and information on the profits of each company collected. It is decided to examine whether the size variables on the register (employment and turnover) are suitable for use in regression estimation of profits.

Unit Number	*Employment*	*Turnover (£m)*	*Profits (£m)*
1	72	68	8
2	92	51	13
3	113	54	12
4	111	226	17
5	136	137	9
6	168	196	23
7	210	261	25
8	212	244	21
9	225	187	19
10	231	289	28
11	264	256	17
12	297	358	28
13	415	399	43
14	432	556	80

A statistician fits three models, in which profits is the response variable, to the data:

with employment as the explanatory variable,

with turnover as the explanatory variable,

with employment and turnover as explanatory variables.

Results are given in the following table ("Emp" is employment, "Tv" is turnover).

Variables in model	R^2 **for model**	**Regression coefficients (Standard error)**		
		Constant	*Emp*	*Tv*
Constant, Emp	0.729	–5.47	0.141	
		(5.91)	(0.025)	
Constant, Tv	0.812	–3.15		0.118
		(4.43)		(0.016)
Constant, Emp, Tv	0.814	–3.99	0.021	0.103
		(5.15)	(0.057)	(0.046)

Using these results, decide upon the most appropriate model of the three.

(8)

(ii) What further steps would you suggest to help make this decision? What are your overall conclusions regarding regression estimation using this sample?

(12)

4. (a) (i) Describe briefly how to construct and use Shewhart charts for the proportion of faulty components being produced on a continuously operating production line.

(8)

(ii) Explain the purposes of cusum charts and how they can be set up.

(6)

(b) When the relationship between an observed variable Y and a predictor x is curved, two possible ways of examining it are (i) to use a polynomial regression, and (ii) to use an equation that can be linearised by transforming one or both of x and Y.

Discuss briefly the advantages and disadvantages of these two methods.

(6)

SOLUTIONS

Question 1

(i) The key assumption is that the experimental errors are independent N(0, σ^2) variables (note constant σ^2).

Totals are

Method 1	Method 2	Method 3
78.6	79.3	86.9

Subj. 1	Subj. 2	Subj. 3	Subj. 4	Subj. 5	Subj. 6	Subj. 7	Subj. 8
23.2	30.8	26.8	26.5	32.9	39.5	27.2	37.9

The grand total is 244.8. $\Sigma\Sigma y_{ij}^2 = 2585.22$.

"Correction factor" is $\frac{244.8^2}{24} = 2496.96$.

Therefore total SS = 2585.22 – 2496.96 = 88.26.

SS for methods = $\frac{78.6^2}{8} + \frac{79.3^2}{8} + \frac{86.9^2}{8} - 2496.96 = 5.30$.

SS for subjects = $\frac{23.2^2}{3} + \frac{30.8^2}{3} + ... + \frac{37.9^2}{3} - 2496.96 = 78.47$.

The residual SS is obtained by subtraction.

Hence the analysis of variance table is as follows.

SOURCE	DF	SS	MS	F value
Methods	2	5.30	2.65	8.26 Compare $F_{2,14}$
Subjects	7	78.47	11.21	34.95 Compare $F_{7,14}$
Residual	14	4.49	0.3207	$= \hat{\sigma}^2$
TOTAL	23	88.26		

Upper critical points of $F_{2,14}$ and $F_{7,14}$ are as follows.

	5%	1%	0.1%
$F_{2,14}$	3.74	6.51	11.78
$F_{7,14}$	2.76	4.28	7.08

Solution continued on next page

The F value for methods is highly significant; we have strong evidence that not all the methods are the same in terms of mean clotting time. The F value for subjects is very highly significant. We have very strong evidence that not all the subjects are the same in this regard; the analysis has detected and removed a large systematic source of variation.

To investigate method differences, we need the method means, which are

Method 1 : 9.825 Method 2 : 9.9125 Method 3 : 10.8625.

The least significant difference between any pair of these means is

$$t_{14}\sqrt{\frac{2\times 0.3207}{8}} = 0.283 t_{14} \quad \text{where} \quad t_{14} = \begin{cases} 2.145 & \text{at } 5\% \\ 2.977 & \text{at } 1\% \\ 4.140 & \text{at } 0.1\% \end{cases}$$

so the least significant differences are 0.607 for 5%, 0.842 for 1% and 1.172 for 0.1%. Clearly methods 1 and 2 do not appear to differ in mean clotting time but there is strong evidence that method 3 has a higher mean clotting time than either of the others.

(ii) In the analysis of variance now, there will be no "subjects" term, only "methods" and "residual". The new residual will include *both* the amount previously classified as residual *and* the amount previously classified as the subjects term. Thus the new analysis of variance table is as follows.

SOURCE	DF	SS	MS	F value
Methods	2	5.30	2.65	0.67 Compare $F_{2,21}$
Residual	21	82.96	3.95	$= \hat{\sigma}^2$
TOTAL	23	88.26		

The F value for methods is now not significant – we have no evidence to reject the null hypothesis that the methods are the same in terms of mean clotting time. This is because the apparent underlying variability in the data is now very high, due to the consistent but unidentified between-subject variation. Blocking therefore greatly increased the power to detect differences between the methods.

Question 2

(i) The number of defectives in a sample, X, has the binomial distribution with parameters 20 and p, i.e. $X \sim \mathrm{B}(20, p)$.

$$P(\text{accept batch} \mid p) = P(X = 0 \text{ or } 1 \mid p) = (1-p)^{20} + 20p(1-p)^{19}$$
$$= (1-p)^{19}(1 + 19p).$$

p = 0.01	:	P(accept batch)	= 0.9831
0.05	:		0.7358
0.1	:		0.3917

(ii) The batch is accepted in the following cases.

Number of defectives in first sample	
0	(Second sample not taken)
1	(Second sample not taken)
2	Second sample has 0 defectives

So P(accept batch)

$= (1-p)^{19}(1 + 19p)$
$+ P$(2 defectives in first sample and 0 defectives in second sample)

$$= (1-p)^{19}(1 + 19p) + \frac{20\times19}{2} p^2(1-p)^{18} \times (1-p)^{20}.$$

The values of this are as follows.

$p = 0.01$: $0.9831 + (0.01586 \times 0.81791) = 0.9831 + 0.01297 = 0.9961$

$p = 0.05$: $0.7358 + (0.18868 \times 0.35849) = 0.7358 + 0.06764 = 0.8034$

$p = 0.1$: $0.3917 + (0.28518 \times 0.12158) = 0.3917 + 0.03467 = 0.4264.$

Solution continued on next page

(iii)

Scheme (i) P(reject batch) = 0.0169 for $p = 0.01$
0.2642 for $p = 0.05$
0.6083 for $p = 0.1$.

Let S = total number inspected. $E(S) = 20P(\text{accept batch}) + 1000P(\text{reject batch})$.

The values of $E(S)$ are as follows.

$p = 0.01$: $(20\times0.9831) + (1000\times0.0169) = 36.6$

$p = 0.05$: $(20\times0.7358) + (1000\times0.2642) = 278.9$

$p = 0.1$: $(20\times0.3917) + (1000\times0.6083) = 616.1$.

Scheme (ii)

$E(S) = 20P(\text{accept batch based on first sample})$
$+ 40P(\text{2 defectives in first sample and 0 in second})$.

The values of $E(S)$ are as follows.

$p = 0.01$: $(20\times0.9831) + (40\times0.01297) + (1000\times0.0039) = 24.1$

$p = 0.05$: $(20\times0.7358) + (40\times0.06764) + (1000\times0.1966) = 214.0$

$p = 0.1$: $(20\times0.3917) + (40\times0.03467) + (1000\times0.5736) = 582.8$.

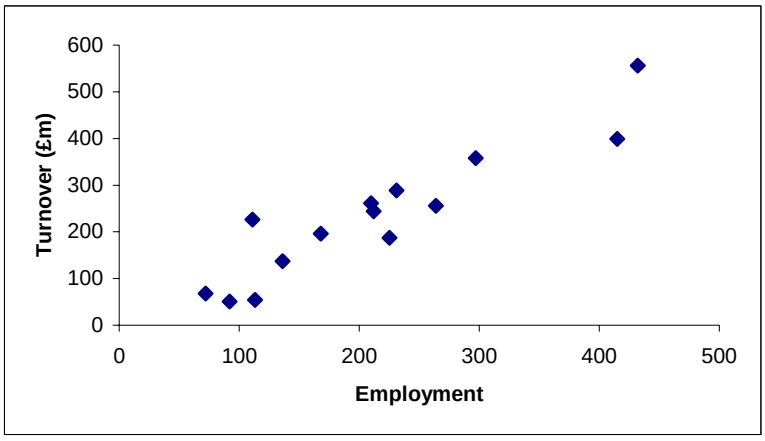
Scheme (ii) will have lower inspection cost than scheme (i) for these values of p.

Question 3

(i) For both one-variable models, the standard errors quickly show that the coefficient of the explanatory variable is significantly different from zero but the constant is not. We can compare these models in terms of their values of R^2: it appears that the second model is the better. R^2 for this model is virtually the same as for the full (two-variable) model, so it seems that there is little if anything to gain by using the two-variable model. Further, that model is a little difficult to interpret: the SEs show that, considered on their own, neither the constant nor the coefficient of employment is significantly different from zero, and the significance of the coefficient of turnover is fairly marginal.

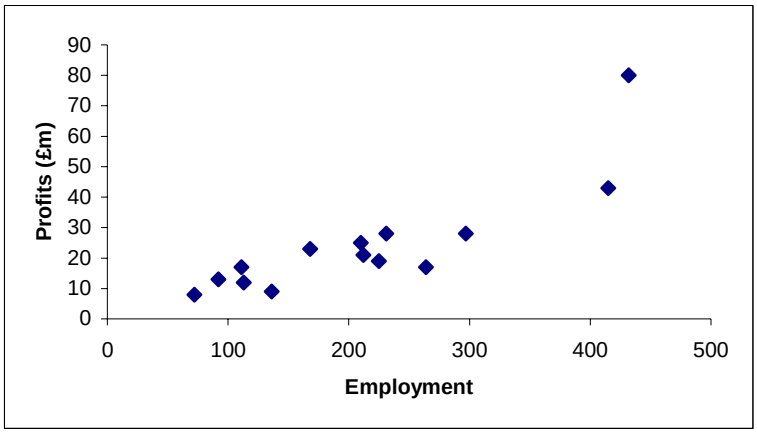
Overall, the one-variable model using turnover seems the most appropriate, judged on these results.

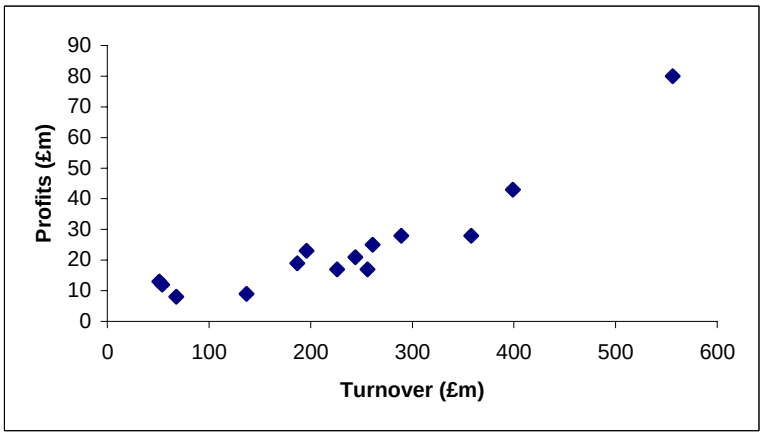
(ii) Scatter diagrams for all pairs of variables will be informative. The diagram for employment and turnover shows a fairly strong linear relationship between the two potential explanatory variables, and this suggests that only one of them is needed in the model.



Solution continued on next page

The relationship between profits and employment is somewhat less satisfactorily linear than that between profits and turnover (though, on a purely visual basis, there is not very much to choose between them):





Overall, then, the best model seems to be $y = -3.15 + 0.118x_2$, where y is profit in £m and x_2 is turnover in £m. It explains 81.2% of the total variation in y and thus appears to be a reasonable model. The fit obtained by using employment instead of turnover is less good. Nor does it seem worthwhile to include employment as a second explanatory variable as well as turnover. Including them both leads to less precise fitting, as is shown by the larger standard errors of the regression coefficients. However, if there are other data available for variables not used so far, these could be explored as possible additional explanatory variables in a larger model.

Question 4

Part (a)(i)

The basic principles are the same as those for a control chart of means. Individual items are inspected and classified as either "sound" or "faulty" (other words such as "acceptable" and "defective" are of course also commonly used). Sample sizes have to be much larger than for means since each individual item yields very little information.

Let $\hat{p} = r/n$ be the proportion found faulty. The sample size n is usually large enough for the underlying distribution of this to be approximated by

$$\mathrm{N}\left(\hat{p}, \frac{\hat{p}(1-\hat{p})}{n}\right).$$

For this to be satisfactory, n needs to be at least 50; and if π is the true proportion faulty, the rule $n\pi \geq 10$ avoids serious problems when π is very small. Thus n may be as much as 200, or even more, per sampling.

On the chart, a horizontal line is drawn at the target value of π, and the successive observations of r/n are plotted. Upper and lower warning and action lines are marked at

$\pi \pm 3.29\sqrt{\dfrac{\pi(1-\pi)}{n}}$ for 99.9% limits (3 might be used as an approximation to 3.29)

and

$\pi \pm 1.96\sqrt{\dfrac{\pi(1-\pi)}{n}}$ for 95% limits (2 might be used as an approximation to 1.96),

where π denotes the target value. (Other limits are sometimes used.) Lower limits calculated to be negative are set at zero.

In some situations there will not be a target value of π, in which case a value of $\hat{p}$ based on a large quantity of historic data would be used.

A typical rule would be to stop the process for examination when a plotted value of r/n goes above the upper action line or when two successive values are above the upper warning line. Values below lower lines are not usually serious, though they may be taken as evidence of excessive variability in the process.

Solution continued on next page

Part (a)(ii)

If it is particularly important to detect any changes that may occur in a process mean value, a cumulative sum (cusum) chart is useful; it also indicates the time point at which a change occurred.

Successive sample means $\bar{x}_1, \bar{x}_2, \ldots$ are used to form a series of "cusums"

$$s_1 = (\bar{x}_1 - k), \quad s_2 = \sum_{i=1}^{2} (\bar{x}_i - k), \quad \ldots, \quad s_r = \sum_{i=1}^{r} (\bar{x}_i - k), \quad \ldots$$

where k is a reference value, either specified as a target or calculated from past data and used as though it were the population mean. These are plotted on a chart. If the process is in control, the sample means will remain fairly close to k with some positive and some negative differences, and the cusum chart will be roughly horizontal. If an increase in process mean occurs, most differences will then be positive and the cusum chart will show a steady increase; similarly, a steady decrease will be shown if the process mean decreases. Cusum charts are sensitive in detecting such changes, as in effect they contain information on all the past history of the process; it is usually possible to pinpoint the time of the change quite accurately. Nevertheless, suitable scales have to be developed and used to interpret charts consistently; some methods are proposed in text books.

In order to decide whether a major change has occurred, a V-mask can be helpful, being similar in purpose to the limit lines on a Shewhart chart. Details of these are given in text books, and there are alternatives according to what change is expected if the system goes wrong.

Part (b)

If the model $Y = \alpha + \beta x$ is extended to $Y = \alpha + \beta x + \gamma x^2$, the resulting parabolic function can be useful if Y appears to have a maximum or minimum somewhere in the range of x for which data are available, and more generally can help to explain relationships that are curved within the region where data are available. When the fit is still not very good, there is a temptation to go on adding further terms, cubic, quartic, and so on – but interpreting the meaning of the coefficients of these powers of x can be very difficult or unconvincing (or both). However, Normality of the residuals may still be assumed.

There may be a more "natural" model that can be linearised: for example, $Y = ae^{bx}$ becomes $\log Y = \log a + bx$. Such models are, for example, quite often found in biological work where a response is not "additive" but "multiplicative". Other power-law relationships also occur in biology and agriculture, and elsewhere; for example, $Y = cx^k$ can be transformed to $\log Y = \log c + k \log x$; here, $\log c$ and k are the parameters to be estimated. It should not be too difficult to explain parameters in an exponential or power model; a better understanding of the process leading to the data may be possible. **BUT** are the residuals still Normally distributed? If we assumed they were Normal before transforming, can they still be so assumed afterwards?