THE ROYAL STATISTICAL SOCIETY

2003 EXAMINATIONS – SOLUTIONS

HIGHER CERTIFICATE

PAPER III
STATISTICAL APPLICATIONS AND PRACTICE

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.

Higher Certificate, Paper III, 2003. Question 1

Summary:

Dial type	1	2	3	
Total	276	327	294	
Number of tests n_i	8	6	7	Total 21
Mean number of errors $\bar{x}_i$	34.5	54.5	42.0	

(i) Analysis of variance

Source of variation	df	SS	Mean Square	*F* ratio
Dial type	2	1377	688.50	6.67
Residual (Error)	18	1858	103.22	
Total	20	3234		

The F ratio of 6.67 is referred to $F_{2,18}$ and is very highly significant (p = 0.007). We reject the null hypothesis that the mean numbers of errors with the three dial types are all the same. We deduce that at least one mean is different from the other two.

We have assumed that all sets of data come from Normally distributed populations with the same variance σ^2.

(ii) We have $\bar{x}_2 - \bar{x}_1 = 20.0$, and the standard error of this estimate is $\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}$ = $\sqrt{103.22\left(\frac{1}{8}+\frac{1}{6}\right)}$ = 5.487. [Thus the difference between the means for dial types 1 and 2 is very highly significant: test statistic is 20.0/5.487 = 3.645, refer to t_{18}.] The 5% critical value for t_{18} is 2.101, so 95% confidence limits for the true population mean difference $\mu_2 - \mu_1$ are

$$20.0 \pm (2.101 \times 5.487) \quad \text{or} \quad 20.0 \pm 11.53, \quad \text{i.e. } (8.47, 31.53).$$

This means that, on the basis of these experimental data, we can say with 95% confidence of being correct that the calculated interval does contain the true value of $\mu_2 - \mu_1$.

(iii) Residuals could be calculated for the 21 observations, and their pattern studied either as a Normal probability plot or by plotting residuals against fitted values.

The variances within the three dial types could be checked for equality, but no good, sensitive, test exists for small amounts of data such as we have here.

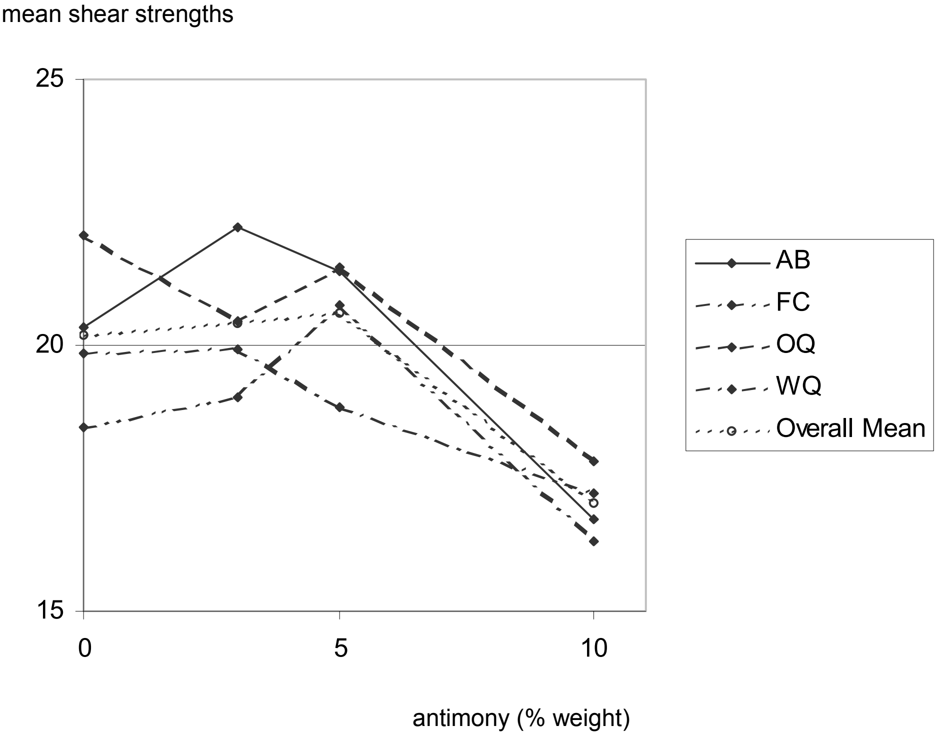
Outliers, if any, could be checked for possible recording error or change in background conditions. Any outliers could be removed from the data before re-doing an analysis.

Sometimes a transformation (such as log) will make data behave more like data from Normal homoscedastic distributions.

Higher Certificate, Paper III, 2003. Question 2

(i) If two (or more) "factors", qualitative or quantitative, are included in the same experiment at various levels, it is often the case that the response to one factor depends on the level at which the other factor is applied. For example, in this experiment, the response to a given % antimony may be different according to which cooling method has been used for a particular experimental unit. When this happens, such factors are said to *interact*.

(ii) The row in the analysis of variance which refers to the possible interaction contains the p-value 0.152, which means that there is no real evidence of interaction (since $p > 0.05$). A graph of the means for the different cooling methods can help to illustrate this (the overall mean is also shown):-



(NB: For clarity in the diagram, and to avoid taking up excessive space, the vertical axis is shown with a "false origin" at 15.)

The graph shows some departure from "parallelism" in the patterns for the cooling methods, but this is not great when compared with (residual) natural variation.

(iii) Since there is no interaction, the main effects of % antimony and cooling method can be studied directly. Both are significant. The main characteristic for % antimony is the drop at 10% compared with all the others. This is very large and is the reason why "$p = 0.000$". For cooling, AB and OQ give higher strengths than FC and WQ. Cooling has $p = 0.004$, still clearly highly significant though the difference is not so big as that given by the drop for 10% antimony.

As usual, the variances underlying all sets of data are assumed to be the same, and the residual term in the appropriate linear model is assumed to be Normally distributed.

Higher Certificate, Paper III, 2003. Question 3

(i) The likelihood of the sample of data is $L = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$ = $e^{-n\lambda}\lambda^{\Sigma x_i} / \Pi x_i!$

$\therefore \log L = -n\lambda + (\Sigma x_i)\log\lambda - \Sigma\log(x_i!)$.

$\frac{d}{d\lambda}(\log L) = -n + \frac{\Sigma x_i}{\lambda}$. Solving $\frac{d}{d\lambda}(\log L) = 0$ gives $\hat{\lambda} = \bar{x}$.

$\frac{d^2}{d\lambda^2}(\log L) = -\frac{\Sigma x_i}{\lambda^2}$ which is < 0 since all $x_i > 0$ (and so $\hat{\lambda} > 0$). Hence this gives a maximum.

(ii) (a) If a Poisson distribution gives a good fit, its mean is estimated by the sample mean $\bar{x}$, since we have no more specific information about λ.

$$\bar{x} = \frac{\Sigma fx}{\Sigma f} = \frac{(0\times18)+(1\times25)+(2\times13)+(3\times10)+(4\times6)+(5\times3)}{75} = \frac{120}{75} = 1.6.$$

Probabilities expected with $\lambda \equiv \hat{\lambda} = \bar{x}$ are $e^{-1.6}(1.6)^x / x!$ for x = 0, 1, 2, … . Expected frequencies are 75 times these.

$P(0) = e^{-1.6} = 0.2019$ $P(1) = 1.6e^{-1.6} = 0.3230$ $P(2) = 0.2584$

$P(3) = 0.1378$ $P(\geq 4) = 0.0788$.

(Each quoted probability is accurate to 4 decimal places.)

Hence we have

x	0	1	2	3	≥ 4	Total
Observed frequency	18	25	13	10	9	75
Expected frequency	15.14	24.23	19.38	10.34	5.91	

A chi-squared goodness of fit test has 3 degrees of freedom, since there are 5 categories of data and 1 parameter estimated.

Test statistic = $\frac{(2.86)^2}{15.14} + \frac{(0.77)^2}{24.23} + \frac{(6.38)^2}{19.38} + \frac{(0.34)^2}{10.34} + \frac{(3.09)^2}{5.92}$ = 4.29, which is not significant (the 5% point of χ_1^2 is 7.81). Hence a Poisson distribution is an acceptable model for these data.

Continued on next page

(b) This is a reasonably large sample of data, although the mean value of 1.6 is somewhat low for using a Normal-approximation confidence interval. An approximate 95% confidence interval for λ is

$$\hat{\lambda} \pm 1.96\sqrt{\hat{\lambda}/75}\ , \quad \text{i.e. } 1.6 \pm 0.286 \quad \text{or} \quad (1.31, 1.89).$$

[If the Poisson distribution is NOT assumed, use $\Sigma fx^2 = 338$, giving $s^2 = \frac{1}{74}\left(338 - \frac{(120)^2}{75}\right) = 1.9730$, so $s = 1.4046$. Thus the interval is $1.6 \pm \left(1.96 \times 1.4046/\sqrt{75}\right) = 1.6 \pm 0.318$, i.e. (1.28, 1.92).]

Higher Certificate, Paper III, 2003. Question 4

(i) Exponential smoothing forecasts $\hat{x}_t$ as $\hat{x}_{t-1} + \alpha(x_{t-1} - \hat{x}_{t-1})$, i.e.

$$\hat{x}_t = \alpha x_{t-1} + (1-\alpha)\hat{x}_{t-1}.$$

α is a value between 0 and 1, combining the previous forecast $\hat{x}_{t-1}$ and the actual observed x_{t-1} as a weighted average.

For 2000, we use $\hat{x}_{1999}$ and x_{1999} :

$$\hat{x}_{2000} = (0.8)(19863) + (0.2)(19236) = 19737.6 \approx 19738.$$

Of course all earlier data are represented to some extent in this forecast.

(ii) A high value of α, such as 0.8, is appropriate if there is little previous experience or if there appears to have been some change in pattern of the data which makes older data less relevant.

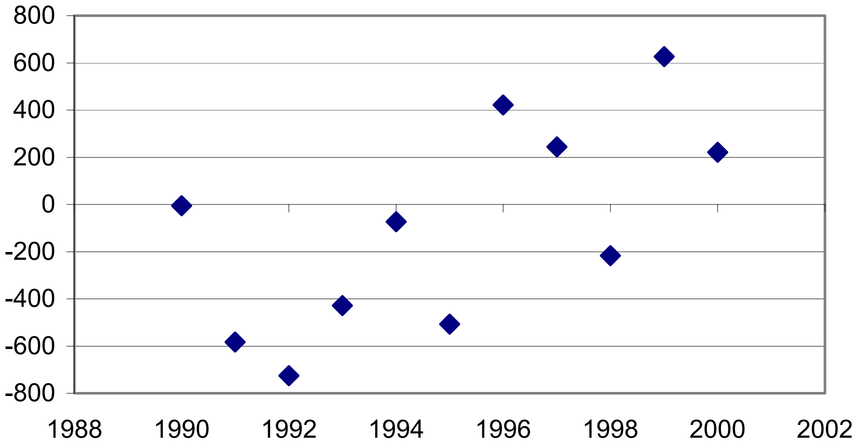
(iii) Mean absolute deviation is a possible method. This is $MAD = \frac{1}{n}\sum_{t=1}^{n}|x_t - \hat{x}_t|$.

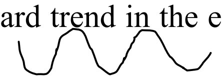
It is a general method, not restricted to exponential smoothing, where various possible models are being compared. The model with minimum *MAD* is usually preferred.

Minimum MSE (mean square error), and some others, have also been suggested; the differences between x_t and $\hat{x}_t$ for past data form the basic criterion.

(iv) $\hat{x}_{2001} = 0.8x_{2000} + 0.2\hat{x}_{2000} = (0.8)(19959) + (0.2)(19738) = 19915.$

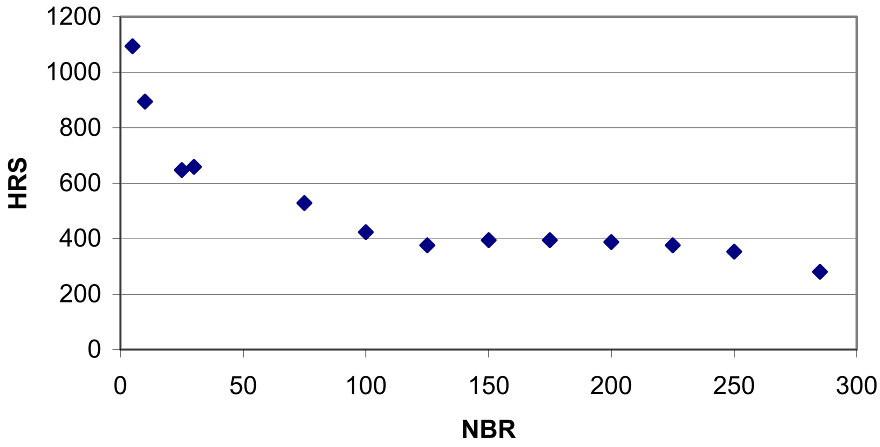
(v) Plot of error (= expenditure – forecast) against time.



Initially, forecasts have proved to be too high; later, they are almost all too low. From 1992 onwards, this method has shown an upward trend in the error. Also there seems to be some tendency to a "cycling" pattern:  … . A slightly more complex model may be needed.

Higher Certificate, Paper III, 2003. Question 5

(i)



The relation is not linear. It began by showing an exponential-type fall over the first 100 or so and then levelled off, with a suggestion of a further curved downward trend after about 200.

(ii) HRS versus logNBR gives the largest percentage of variation explained ($R^2 = 96\%$). (HRS versus 1/NBR is also fairly good, but some way behind this one.)

(iii) HRS = 1306 – 181 logNBR. The constant is a baseline or average cost estimated from these data. The coefficient of logNBR ("log" here implies to base *e*) gives the reduction (in this case) in HRS (thousands of man-hours) for every increase of 1 in logNBR, i.e. every 2.71828 along the number scale. This is an average reduction.

If the residuals were available, we would look at them to see if they were "random" or if they showed some pattern (for example, all the middle ones were of one sign and the first and last were of the other sign). If so, this would suggest that the model could still be improved, even given the high value of R^2.

(iv) Perhaps logHRS against logNBR might show improvement.

Higher Certificate, Paper III, 2003. Question 6

(i) $\hat{p}_A = \frac{84}{150}, \ n_A = 150. \qquad \hat{p}_B = \frac{126}{200}, \ n_B = 200.$

95% limits for the true value of $(p_A - p_B)$ are estimated as

$$(\hat{p}_A - \hat{p}_B) \pm 1.96\sqrt{\frac{\hat{p}_A(1-\hat{p}_A)}{n_A} + \frac{\hat{p}_B(1-\hat{p}_B)}{n_B}},$$

$$\text{i.e. } (0.56 - 0.63) \pm 1.96\sqrt{\frac{0.56 \times 0.44}{150} + \frac{0.63 \times 0.37}{200}}$$

$$= -0.07 \pm 1.96\sqrt{0.002808} = -0.07 \pm 0.104,$$

i.e. (–0.174, 0.034).

The interval contains zero, so we should not claim that one journal is significantly better than the other. However, with 95% confidence, we may claim that the difference between them ranges from 3.4% in favour of A to 17.4% in favour of B.

(ii) The two statements are alternatives, which together with (I) make up the responses of the whole sample. Hence they are not independent, and the test in part (i) assumes that they are.

(iii) Although the statements relate to two different issues of the journal, they are answered by (at least some of) the same people, and so once again the responses do not come from independent samples. The two proportions will most likely be correlated.

(iv) Suppose n is the required sample size. Using $\hat{p}$ = 0.63, as in part (i), Var($\hat{p}$) is estimated as (0.63)(0.37)/n. An approximate value for n is found by making $0.05 = 1.96 \times \text{SE}(\hat{p})$. This gives

$$\left\{\frac{0.05}{1.96}\right\}^2 = \text{Var}(\hat{p}) = \frac{(0.63)(0.37)}{n},$$

$$\text{i.e. } n = (0.63)(0.37)\left\{\frac{1.96}{0.05}\right\}^2 = 358.2,$$

so at least 359 responses are needed.

Because this is a large sample, and p is not too far from ½, this approximation will be satisfactory. Also we are told that there are a very large number of subscribers, so that a "finite population correction" is unnecessary (even if we knew N).

Higher Certificate, Paper III, 2003. Question 7

(i) (a) Non-respondents tend not to be a random part of the whole population, but instead particular types of person are more likely to fail, or refuse, to reply. Their responses, if known, would quite likely be different from other parts of the population. Unless they are represented, the results of the survey will not validly apply to the whole population. Besides introducing this bias, intended sample size is reduced by non-response and so precision suffers.

(b) Some possible procedures are as follows.

- Send the questionnaire again, once or twice more
- Send reminder letters (without the questionnaire)
- Telephone people who have not responded
- Visit those who have not responded, perhaps only a sample of them

These all require identification of non-responders, usually by means of a number on the questionnaire which is kept separate from the answers to preserve anonymity.

(ii) (a) Strategy A will lead to a sample consisting of those who are easiest to locate, so even if the list is constructed in a properly random way the actual members used will not have been selected at random from the list. Any who refuse at first request will be ignored rather than any attempt being made to persuade them. Since there is a team of interviewers, the more efficient of these may carry out a higher proportion of the 600 (more quickly). Or, alternatively, quickly completed interviews may not have been done so thoroughly. Why stop at 600 instead of attempting to get as many as possible of the originally selected list?

Strategy B will also lead to willing and easily available heads of household being selected, so that the first sample of (600 – m) could suffer considerable bias. The second sample of m ought to be more representative, but the quality of the final data will be affected by how large m is (the larger the better to avoid bias). Office work is also increased by this method.

Continued on next page

(b) The number R of respondents will be Binomial(n, ¾). Since n is bound to be a large number, we may use a Normal approximation:

$$R \sim \mathrm{N}\left(\frac{3n}{4}, \frac{3n}{16}\right).$$

Thus, approximately,

$$Z = \frac{R - \frac{3n}{4}}{\sqrt{\frac{3n}{16}}} \sim N(0,1),$$

and the upper 99% point of Z is 2.326.

Hence

$$2.326 \le \frac{600 - \frac{3n}{4}}{\frac{1}{4}\sqrt{3n}}, \qquad \text{i.e.} \qquad 2.326 \le \frac{2400 - 3n}{\sqrt{3n}} \ .$$

Solve this for equality, using only the upper value for n:-

$$2.326\sqrt{3n} = 2400 - 3n, \qquad \text{or} \qquad 3n + 2.326\sqrt{3n} - 2400 = 0.$$

Write $\sqrt{3n} = x$, so we have $x^2 + 2.326x - 2400 = 0$, and the roots are

$$x = \frac{-2.326 \pm \sqrt{(2.326)^2 + 9600}}{2} = \frac{1}{2}(-2.326 \pm 98.0072)$$

$$= -50.1666 \text{ or } +47.8406,$$

giving

$$n = \tfrac{1}{3}x^2 = 838.9 \quad (\text{or } 762.9). \quad \text{Take } n = 839.$$

Higher Certificate, Paper III, 2003. Question 8

Main points which could be made include the following.

> Total number of employees remained fairly steady, around 22 million, but the ratio of males to females decreased from 13.4/9.4 = 1.43 in 1978 (and 1.35 in 1981) to 11.5/11.3 = 1.02 in 1997 (1.07 in 1991).
>
> Percentage of males in category H remained much the same, as did that of females, but the average percentage for males was about 18 and for females about 40 (presumably nursing and medical services are included in H, which would provide an explanation).
>
> Percentage of workers overall (both sexes) in category B fell sharply between 1978/1981 and 1991/1997.
>
> Percentage of workers in category C increased for both sexes between 1978/1981 and 1991/1997.
>
> In D, E and F the percentages over time remained similar, with more males than females.
>
> In G, the number of employees had dropped sharply by 1997.

Some calculations of actual numbers (bottom row × appropriate percentages) would help to emphasise the drop in numbers of workers (both sexes) in manufacturing, and the increases in numbers for financial and business services.

A combination of percentages and actual numbers would indicate a noticeable increase in male employment in category A. For females, numbers increase though not percentages.

Graphs of "time series" for the two sexes and four years, a single graph for each category, would help to show the changes, in categories B and C particularly.

Bar charts could be used to show actual numbers or percentages over time, one for each year. Because of the presence of several categories with small percentages, annual pie-charts would be slightly less easy to appreciate (but quite valid).

Some of the categories are rather broad, and explanations of changes therefore not always possible even for a UK commentator.