# THE ROYAL STATISTICAL SOCIETY

# 2001 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, 2001. Question 1

(i) Using the differences, test the null hypothesis "mean difference = 0", assuming Normality of the distribution of differences.

$\bar{d} = -12.3, \quad s_d^2 = (24.3176)^2$; so test statistic is $\dfrac{-12.3 - 0}{24.3176/\sqrt{10}} = -1.60$, which is not significant as an observation from $t_9$.

(ii) Correction term $= 4831^2 / 20 = 1166928.05$.

$$\text{SS for weeks} = \frac{1}{10}\left(2477^2 + 2354^2\right) - \text{correction} = 1167684.50 - 1166928.05$$
$$= 756.45$$

$$\text{SS for patients} = \frac{1}{2}\left(258^2 + \ldots + 613^2\right) - \text{correction} = 1238009.50 - 1166928.05$$
$$= 71081.45$$

Analysis of Variance

| ITEM | DF | SS | MS | | |
|---|---|---|---|---|---|
| Patients | 9 | 71081.45 | 7897.94 | $F_{9,9} = 26.71$ | sig at 0.1% |
| Weeks | 1 | 756.45 | 756.45 | $F_{1,9} = 2.56$ | not significant |
| Residual | 9 | 2661.05 | 295.67 | | |
| TOTAL | 19 | 74498.95 | | | |

(iii)

| Ranking of \|diff\| | 4 | 5 | 6 | 8 | 9 | 24 | 25 | 26 | 40 | 56 |
|---|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
| Sign | + | + | + | − | − | − | + | − | − | − |

Sum of + ranks is $S_+ = 13$; $S_- = 42$. Tables show that for $n = 10$ and at the 5% level in a two-tail test, the smaller of $S_+$ and $S_-$ should be 8 or less for significance.

(iv) The null hypothesis for (i) and (ii) is as stated in (i). The alternative hypothesis is "mean difference ≠ 0". Normality of the data would be required in (ii), not just of the differences. A dot-plot would in either case cast serious doubt on this assumption. The Wilcoxon test does not require any distributional assumption, only that the + and – rankings are randomly placed in the set. In each case we must not reject the null hypothesis because we do not have any statistically significant test results.

(v) $t_9^2 = F_{1,9}$ .

Higher Certificate, Paper III, 2001. Question 2

Totals are:

| | | $N_0$ | 50 | 100 | |
|---|---|---|---|---|---|
| Depth | 8 cm | 4326 | 5028 | 5727 | 15081 |
| | 12 cm | 4633 | 5437 | 6223 | 16293 |
| | | 8959 | 10465 | 11950 | 31374 |

Correction term $= \dfrac{31374^2}{24} = 41013661.50$.

SS for depths $= \dfrac{1}{12}\left(15081^2 + 16293^2\right) = 41074867.50$.

SS for nitrogen $= \dfrac{1}{8}\left(8959^2 + 10465^2 + 11950^2\right) = 41572800.75$.

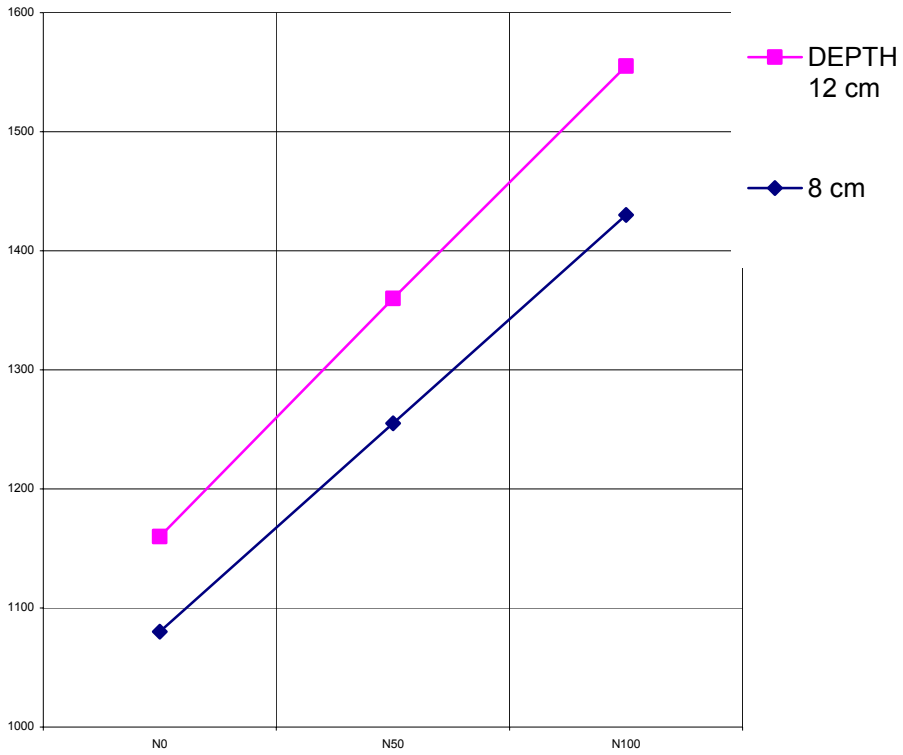
(i) Completed ANOVA is:

| SOURCE | DF | SS | MS |
|---|---|---|---|
| Nitrogen | 2 | 559139.25 | 279569.625 |
| Depth | 1 | 61206.00 | 61206.000 |
| N×D | 2 | 2237.30 | 1118.650 |
| Residual | 18 | 7161.75 | 397.875 |
| TOTAL | 23 | 629744.50 | |

Both the nitrogen effect and the depth effect are obviously very highly significant. Depth 12 gives a yield very significantly greater than depth 8. There is a very significant nitrogen effect.

There is no evidence of interaction ($F_{2,18}$ test statistic is 2.81).

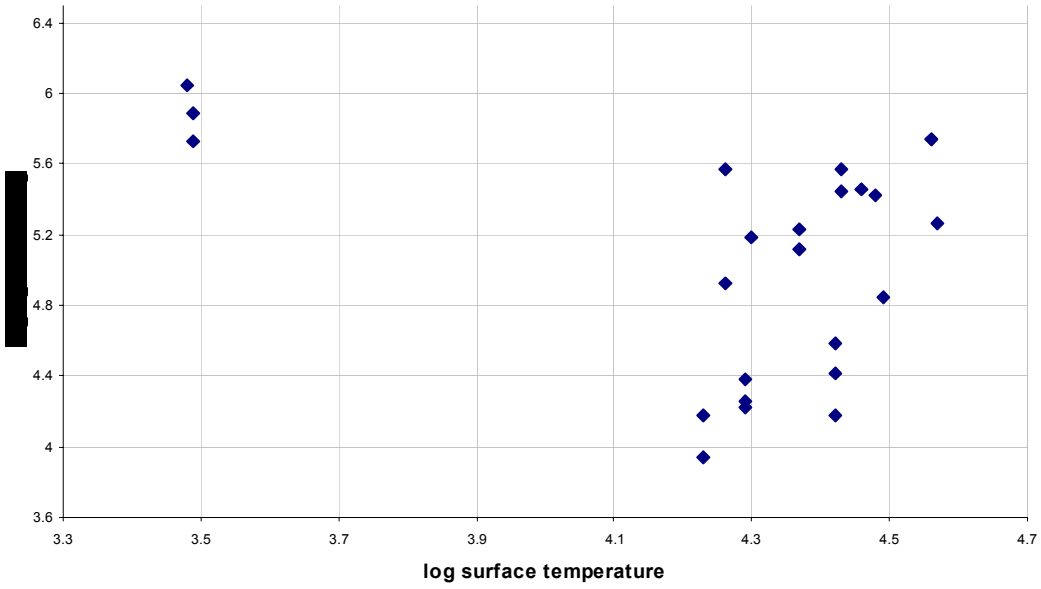
(ii)




(iii) There will be considerably higher yield if 100kg of sulphate of ammonia per acre is used, as compared with 50 or with none at all; and 12cm depth of winter ploughing will give better results than 8cm. (The benefit of greater depth is about the same whichever level of nitrogen is used.)

Higher Certificate, Paper III, 2001. Question 3

(i) The graph below shows that the three points in the top left corner have a large influence in the calculation of slope.




---

(ii) Without the three outlying points, $N = 21$ and $\bar{x} = 4.387$, $\bar{y} = 4.938$.

$$S_{XY} = 455.7101 - \frac{1}{21}(92.13)(103.7) = 0.76339$$

$$S_{XX} = 404.4303 - \frac{92.13^2}{21} = 0.24283$$

Hence $\hat{b} = \frac{S_{XY}}{S_{XX}} = 3.144$.

(iii) This $\hat{b}$ is much larger, and positive, as there is a strong tendency for log $y$ and log $x$ to increase together, except for the three outlying points already mentioned.

To test the null hypothesis $b = 0$ we need the Analysis of Variance:

$$\text{Total SS} = S_{YY} = 519.0608 - \frac{103.7^2}{21} = 6.98032.$$

$$\text{Regression SS} = \frac{S_{XY}^{\ 2}}{S_{XX}} = 2.39989.$$

| SOURCE | DF | SS | MS |
|---|---|---|---|
| Regression | 1 | 2.39989 | |
| Deviations | 19 | 4.58043 | $0.24108 = \hat{\sigma}^2$ |
| TOTAL | 20 | 6.98032 | |

$$\mathrm{Var}\left(\hat{b}\right) = \frac{\hat{\sigma}^2}{S_{XX}} = 0.9928, \quad SE\left(\hat{b}\right) = 0.996 .$$

Value of test statistic is 3.144/0.996 = 3.16. Comparing with $t_{19}$, this is significant at the 1% level. Reject the null hypothesis $b = 0$.

(iv) Including the outliers gives a slope which is negative, but not significantly different from 0; and very little variation (9.9%) is explained. Removing them allows 34.4% of variation to be explained, with a slope that is clearly not zero.

The three points are the highest surface temperature values. They do not seem to be of the same population as the rest; perhaps some different mechanism is operating at high temperatures.

Higher Certificate, Paper III, 2001. Question 4

If using a model based on polynomial trends, long-past observations still have some influence on forecasts. Weighted averages of observations are useful, most recent receiving largest weights. In (i), the forecast $x(\hat{t},1)$ uses weights $\alpha(1-\alpha)^j$ which decay exponentially to 0. The current observation, at $j = 0$, receives most weight. If $\alpha = 1$, the forecast is simply the current observation. This forecasting method only uses at each step one previous value of $x$, one time-unit past, is quick and easy, and is based on a statistical model where the first difference $x_t - x_{t-1}$ is MA(1) in the errors.

(i) Using $\hat{x}(t,1) = \alpha x_t + \alpha(1-\alpha)x_{t-1} + \alpha(1-\alpha)^2 x_{t-2} + ...$

we have

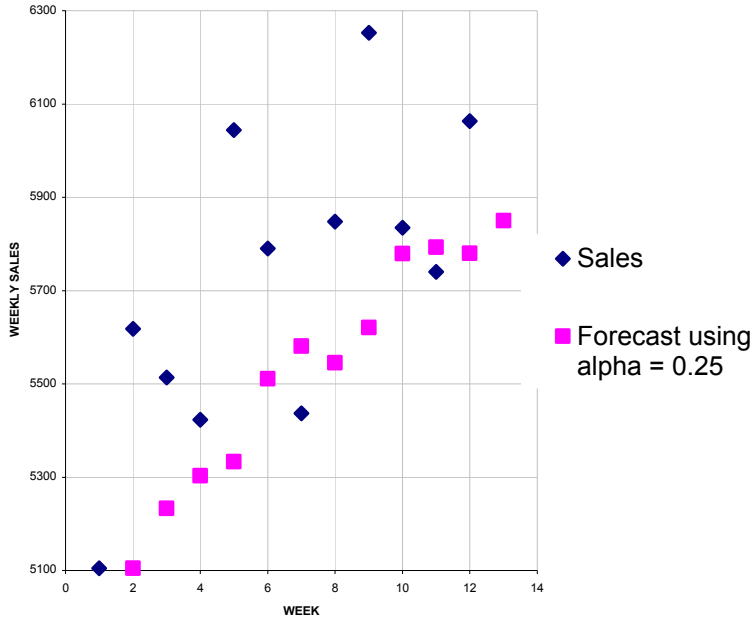
$$\hat{x}(t-1,1) = \alpha x_{t-1} + \alpha(1-\alpha)x_{t-2} + \alpha(1-\alpha)^2 x_{t-3} + ... \ .$$

$$\therefore \ (1-\alpha)\hat{x}(t-1,1) = \alpha(1-\alpha)x_{t-1} + \alpha(1-\alpha)^2 x_{t-2} + \alpha(1-\alpha)^3 x_{t-3} + ...$$

which gives

$$\begin{aligned}\hat{x}(t,1) &= \alpha x_t + (1-\alpha)\hat{x}(t-1,1) \\ &= \alpha\{x_t - \hat{x}(t-1,1)\} + \hat{x}(t-1,1).\end{aligned}$$

(ii) (a)




---

(b) With $\alpha = 0.25$, the forecast for week 3 sales is ¼(5618 – 5105) + 5105, i.e. 5233; then for week 4 the forecast is ¼(5514 – 5233) + 5233, etc. Hence:

| week | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|
| | 5105 | 5233 | 5303 | 5333 | 5511 | 5581 | 5545 | 5621 |

| week | 10 | 11 | 12 | 13 |
|---|---|---|---|---|
| | 5779 | 5793 | 5780 | 5850 |

(c) Forecasts do not seem adequate as they frequently underestimate sales by a large amount. Perhaps using a larger value of $\alpha$, to reduce the weight of past history, would improve this model.

(d) Considering $x_t - \hat{x}(t-1,1)$ as an error, and comparing its average size on two models, would be a guide: we could use a root-mean-square $\left[\dfrac{\sum\{x_t - \hat{x}(t-1,1)\}^2}{11}\right]^{\frac{1}{2}}$ for weeks 2 to 12.

Higher Certificate, Paper III, 2001. Question 5

(i) Mean = $np$; variance = $np(1-p)$.

(ii) $$\text{Mean} = \frac{(0\times 75)+(1\times 44)+(2\times 36)+(3\times 7)+(4\times 4)+(5\times 3)}{169} = \frac{168}{169}$$

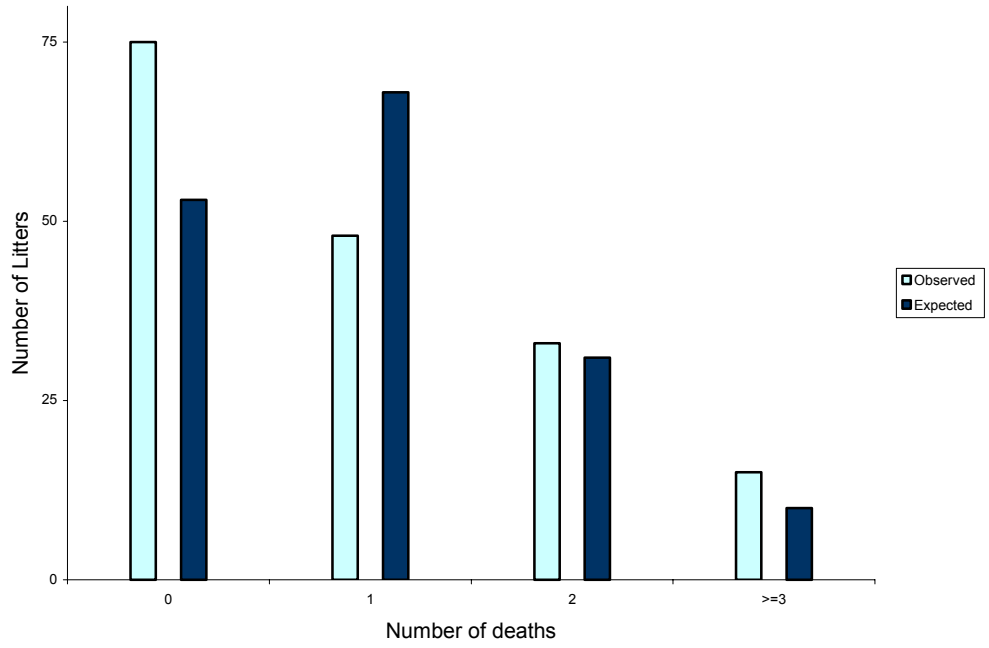
$$= 0.9941 = n\hat{p}.$$

Hence $\hat{p} = \frac{0.9941}{5} = 0.1988, \qquad n\hat{p}(1-\hat{p}) = 0.7965$.

(iii) $p(0) = (1-p)^5 = 0.33014; \quad p(1) = 5p(1-p)^4 = 0.40959$

$p(2) = 10p^2(1-p)^3 = 0.20326; \quad p(\geq 3) = 0.05701.$

Multiplying these by 169 gives the expected frequencies.



| Frequencies | 0 | 1 | 2 | ≥3 | TOTAL |
|---|---|---|---|---|---|
| Observed | 75 | 44 | 36 | 14 | 169 |
| Expected | 55.79 | 69.22 | 34.35 | 9.63 | 168.99 |

$X^2 = \sum \frac{(OBS - EXP)^2}{EXP} = 6.615 + 9.189 + 0.079 + 1.983 = 17.87$ which is highly significant as an observation from $\chi^2_2$ (2 degrees of freedom since an estimated value of $p$ is used).

The null hypothesis of a binomial model is rejected.

(iv) The value of $p$ is assumed the same for each individual in each litter, all independently of one another. The calculated variance $s^2$ = 1.3273 is larger than $n\hat{p}(1-\hat{p}) = 0.7965$. The data observed have more 0s than expected on this model, and fewer 1s; also more of the higher number 3, 4, 5. There is evidence of "overdispersion", the independence and constancy assumptions breaking down.

(v) Mean would be 4.006, new $p$ = old$(1 - p)$, variance the same, histograms the mirror images of old ones. $\chi^2$ same, so inferences same.

Higher Certificate, Paper III, 2001. Question 6

(i) $S(t_0) = P(T \geq t_0) = 1 - \int_0^{t_0} \lambda e^{-\lambda t} dt = 1 - \left[ -e^{-\lambda t} \right]_0^{t_0} = 1 - \left[ -e^{-\lambda t_0} + 1 \right] = e^{-\lambda t_0}$.

ALTERNATIVELY, $\int_{t_0}^{\infty} f(t)\,dt$ may be calculated directly.

(ii) $n = 12, \quad \sum_{i=1}^{12} t_i = 6028$.

$L = \prod_{i=1}^{12} f(t_i) = \lambda^{12} e^{-\lambda \sum t_i}$.

$\ln L \equiv l = 12 \ln \lambda - \lambda \sum t_i$, so $\frac{dl}{d\lambda} = \frac{12}{\lambda} - \sum t_i$

and this is 0 when $\hat{\lambda} = \frac{12}{\sum t_i} = \frac{12}{6028} = 1.9907 \times 10^{-3}$.

$\mathrm{Var}(\hat{\lambda}) \approx \frac{1}{-\mathrm{E}\left( \frac{d^2 l}{d\lambda^2} \right)}$ and $\frac{d^2 l}{d\lambda^2} = -\frac{12}{\lambda^2}$, so this is $\frac{\lambda^2}{12} = 3.302 \times 10^{-7}$.

(iii) Measured in years, $\sum t_i = \frac{6028}{365}$, so $\hat{\lambda} = 0.7266$.

Setting $t_0 = 1$ in (i), $S(1) = e^{-0.7266} = 0.484$.

Higher Certificate, Paper III, 2001. Question 7

(a) (i) Non-response is a problem introduced by people refusing to reply to a survey (or being genuinely unavailable), because they may be different in some ways from those who do reply. People may not have the information to reply to questions, may have different working or leisure habits, may be away more frequently, may live in shared accommodation which is less easy to locate, may be of one particular age-group.

For example, those who play sports would be more likely to be out evenings or weekends, but their views on facilities would be different from those who do not.

(ii) (1) Pilot testing of questionnaires to check clarity, understandability, avoid giving offence by wording or by including sensitive questions.

(2) Use interviewers who are well-trained, understand the aim and purpose of the survey, and the meaning of questions, and are able to put people at their ease.

(3) Give advance notice where appropriate, e.g. to the area or group of people being surveyed.

(4) Revisit those unable to be interviewed through absence.

(b) Light engineering: $n = 125,\ P(\text{improve}) = \dfrac{67}{125} = 0.536$.

Banking & finance: $n = 200,\ P(\text{improve}) = \dfrac{126}{200} = 0.630$.

(i) Using a Normal approximation $\left( p, \dfrac{p(1-p)}{n} \right)$ for each proportion, the variance of their difference is $\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2} = 3.1551 \times 10^{-3}$.

An approximate 95% confidence interval for the true difference $(\pi_2 - \pi_1)$ is $(p_2 - p_1) \pm 1.96\sqrt{3.1551 \times 10^{-3}}$, i.e. $(-0.016,\ 0.204)$.

(ii) Strictly we cannot say they are different because this interval includes 0; but the lower limit is only just below 0 so we might investigate further, if possible, using larger samples.

Higher Certificate, Paper III, 2001. Question 8

(i) The data should be Normally distributed about the respective treatment means; and the variances of all the observations should be the same. The constant variance condition is more important. For these data, the variances within the different treatments are so different that this condition cannot be assumed. There is an obvious relation between mean and variance, which a transformation may be able to correct for.

(ii) $\sqrt{x}$ : rate 3, variance = 13.49.

$\ln x$ : rate 2, mean = 4.827, variance = 0.1212.

$1/x$ : rate 1, variance = 0.0004721;
rate 4, mean = 0.00147.

The logarithmic transformation should be used, since it achieves approximately the same variance for each rate and there is no evidence of a mean–variance relation.

(iii)

| SOURCE | DF | SS | MS | |
|---|---|---|---|---|
| Rates | 3 | 21.153 | 7.051 | $F_{3,8} = 45.6$ |
| Residual | 8 | 1.236 | $0.1545 = s^2$ | |
| TOTAL | 11 | 22.389 | | |

Comparing 45.6 with $F_{3,8}$, there is strong evidence of a difference among the means for the various rates.

(iv) Rate 3: mean $\bar{x} = 5.716$, pooled variance from ANOVA = 0.1545 (8df).

Two-tailed 5% point of $t_8$ is 2.571. $r = 3$. Limits are $\bar{x} \pm t\sqrt{\dfrac{s^2}{r}}$.

This gives $5.716 \pm 2.571\sqrt{\dfrac{0.1545}{3}} = 5.716 \pm 0.583$ i.e. 5.133 to 6.299.

$e^{5.133} = 169.5$; $e^{6.299} = 544.0$ [Note : $e^{5.716} = 303.7$]