THE ROYAL STATISTICAL SOCIETY

2002 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, 2002. Question 1

(i) The main effect terms need to be calculated; the remaining information then follows using what is given, once the degrees of freedom have been completed.

TOTALS:	Foetus	1	167.9	Observer	1	141.0	$N = 36$
		2	236.3		2	137.6	Grand total $G = 558.1$
		3	153.9		3	138.2	
					4	141.3	$G^2/N = 8652.1003$

Corrected $SS_{FOETUS} = \frac{1}{12}\left(167.9^2 + 236.3^2 + 153.9^2\right) - \frac{G^2}{N} = 324.0089$.

Corrected $SS_{OBSERVER} = \frac{1}{9}\left(141.0^2 + 137.6^2 + 138.2^2 + 141.3^2\right) - \frac{G^2}{N} = 1.1986$.

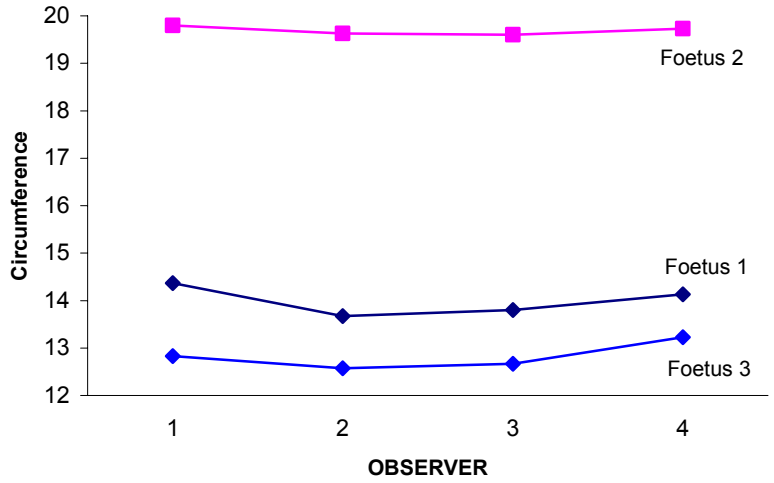
Source	df	Sum of Squares	Mean Square	*F* value
Foetuses (F)	2	324.009	162.004	2112 (very highly sig)
Observers (O)	3	1.199	0.400	5.22 (highly sig)
Interaction (O × F)	6	0.562	0.094	1.22 (not sig).
Error (Residual)	24	1.840	0.0767	
Total	35	327.610		

The interaction term is not significant. Each main effect is highly significant.

(ii) The means are:

		Observer			
		1	2	3	4
Foetus	1	14.37	13.67	13.80	14.13
	2	19.80	19.63	19.60	19.73
	3	12.83	12.57	12.67	13.23

The diagram shows the very large difference between foetuses, the small difference between observers by comparison (even though it is significant at 1%) and the negligible interaction (as the three lines are roughly parallel).



(iii) The four observers did not produce exactly the same results on each foetus, but the differences among observers were small by comparison with those between foetuses. There was no evidence that different observers were measuring the different foetuses inconsistently (i.e. there was no "interaction" between *O* and *F*).

Higher Certificate, Paper III, 2002. Question 2

(a) (i) On the null hypothesis that males and females have equal mean expenditures $(\mu_M = \mu_F)$, against the alternative hypothesis that they do not, and with large enough sample sizes to assume that the difference $(\bar{X}_M - \bar{X}_F)$ between the observed means is approximately Normally distributed, an appropriate test uses $Z = \dfrac{\bar{X}_M - \bar{X}_F}{SE(\bar{X}_M - \bar{X}_F)}$. The estimated variances of the two means are $\dfrac{s_M^2}{n_M}$ and $\dfrac{s_F^2}{n_F}$, and so $SE(\bar{X}_M - \bar{X}_F) = \sqrt{\dfrac{s_M^2}{n_M} + \dfrac{s_F^2}{n_F}}$.

$$s_M^2 = \frac{1}{86}\left(32234.71 - \frac{1098.60^2}{87}\right) = \frac{18362.044}{86} = 213.512.$$

$$s_F^2 = \frac{1}{62}\left(25810.04 - \frac{887.75^2}{63}\right) = \frac{13300.515}{62} = 214.524.$$

$$\bar{x}_M = \frac{1098.60}{87} = 12.6276; \quad \bar{x}_F = \frac{887.75}{63} = 14.0913; \quad \bar{x}_M - \bar{x}_F = -1.464.$$

$$SE(\bar{X}_M - \bar{X}_F) = \sqrt{\frac{213.512}{87} + \frac{214.524}{63}} = \sqrt{2.4542 + 3.4051} = 2.421.$$

Hence the value of Z is $-\dfrac{1.464}{2.421} = -0.605$, which (compare with N(0,1)) is not significant.

There is no evidence of a difference between μ_M and μ_F.

(ii) The assumptions stated in (i) are all that are theoretically necessary. The underlying populations do not need to be Normally distributed nor to have equal variances. The samples are assumed random. The main practical doubt about validity is the existence of zeros in the data. It would be best to base the test on the non-zero items, and additionally compare the proportions of zeros in the two samples.

Continued on next page

(b) (i) The null hypothesis will be $\mu_X = 0$, and we assume X follows a Normal distribution. $n = 15$, $\bar{x} = \frac{32.6}{15} = 2.173$.

$$s^2 = \frac{1}{14}\left(84.18 - \frac{32.6^2}{15}\right) = 0.9521.$$

Test statistic is $\frac{\bar{x} - 0}{\sqrt{\frac{0.9521}{15}}}$ = 8.63, which we refer to t_{14} – very highly significant.

This is very strong evidence against the null hypothesis, which we shall reject.

(ii) The commentator's result is not significant and the null hypothesis "$\mu_1 = \mu_2$" cannot be rejected (μ_i is the mean in year i).

(iii) There is substantial systematic variation from company to company: if x_1 is below average, so is x_2 in most cases. If this between-company variation is removed, by using the differences $x = x_1 - x_2$, the values x should (on the null hypothesis) represent only random variation and give a valid basis for comparison. Clearly in this case the between-company variation was very large, and removing it gave a much more precise comparison of the two years.

Higher Certificate, Paper III, 2002. Question 3

$n = 11$, $\sum t_i = 0$, $\sum t_i^2 = 110$, $\sum y_i = 7751.2$, $\sum y_i^2 = 5491108.76$, $\sum t_i y_i = 1706.3$.

(i) $S_{tt} = 110$, $S_{ty} = 1706.3$ (since $\sum t = 0$); hence $\hat{\beta} = \dfrac{1706.3}{110} = 15.5118.$

$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{t} = \bar{y} = 704.6545$.

$S_{yy} = 5491108.76 - \dfrac{7751.2^2}{11} = 29190.4473$.

$r^2 = \dfrac{S_{ty}^2}{S_{tt} S_{yy}} = 0.9067.$

The sum of squares due to fitting the regression line is $\dfrac{S_{ty}^2}{S_{tt}} = 26467.8154$ and the residual SS is therefore $S_{yy} - \dfrac{S_{ty}^2}{S_{tt}} = 2722.6319$.

This has 11 – 2 = 9 df and so $\hat{\sigma}^2 = 302.5147.$

Thus estimated variances of $\hat{\alpha}$ and $\hat{\beta}$ are given by

$$\mathrm{Var}\left(\hat{\beta}\right) = \frac{\hat{\sigma}^2}{S_{tt}} = 2.7501, \qquad \mathrm{Var}\left(\hat{\alpha}\right) = \hat{\sigma}^2 \left\{\frac{1}{n} + \frac{\bar{t}^2}{S_{tt}}\right\} = \frac{\hat{\sigma}^2}{11} \text{ (since } \bar{t} = 0) = 27.5013.$$

Hence $SE(\hat{\alpha}) = 5.24$ and $SE\left(\hat{\beta}\right) = 1.658.$

(ii) The $\{\varepsilon_i\}$ in the linear model are usually assumed to be independent, from the same Normal distribution, mean 0, variance σ^2. As y increases with t, it is possible that the variance also increases (heteroscedasticity); and due to likely trade-cycles and also non-linearity of the model, independence may also be doubtful.

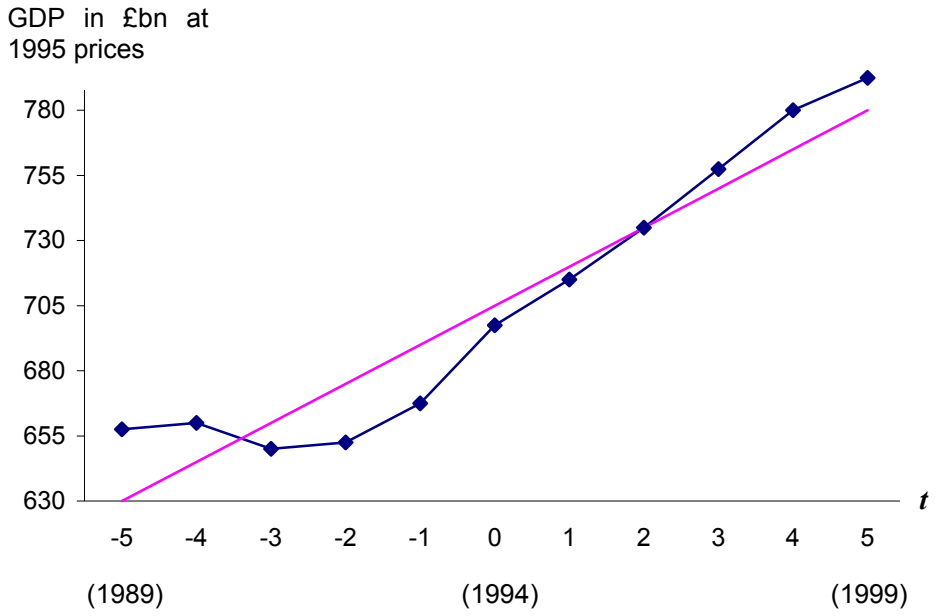
(iii) The correlation coefficient $r = \sqrt{0.9067} = 0.952$. This shows a strong linear relation between y and time. Alternatively, r^2 shows that 90.7% of the variation in the annual GDP figures can be explained as a linear trend in time, with positive slope ($\hat{\beta} = 15.52$). The slope is the average annual increase in GDP over the period, which is £15.52 bn (expressed at 1995 prices); this is over 2% of the annual figure.

$\hat{\alpha} = 704.65$ was the figure estimated for $t = 0$, i.e. 1994, on the basis of all the data (the actual figure allowing for the random variation was 694.6).

Continued on next page

(iv) $t = -3$; $\hat{y} = 704.6545 - (3 \times 15.5118) = 658.12$.

$t = +3$; $\hat{y} = 704.6545 + (3 \times 15.5118) = 751.19$.



(v) Note that because of the "flat" period in the first 4 years (or the drop back in 1991, however this is interpreted), the linear relation is forced to have a somewhat smaller slope than seems necessary to explain 1991 – 1999 well.

For 2000, $\hat{y} = 704.6545 + (6 \times 15.5118)$ = 797.73. Because of the slope not reflecting the years' GDP in late 1990s, this estimate is probably low (but not unreasonable).

For 2010, $\hat{y} = 704.6545 + (16 \times 15.5118)$ = 952.84, but this can only be a guess as the relationship may become curved, or the slope change, or a discontinuity happen (as in 1990/1), and extrapolation so far ahead is not really of any use.

Higher Certificate, Paper III, 2002. Question 4

(i) $A - T$ (Actual – Trend) figures:

		Quarter				
		1	2	3	4	
Year	1996	.	.	95.500	204.375	
	1997	–283.500	–105.875	200.875	253.625	
	1998	–242.500	–168.125	101.500	265.125	
	1999	–236.500	–149.750	218.625	77.735	
	2000	–269.750	–78.500	.	.	
Sum		–1032.250	–502.250	616.500	800.860	
Mean		–258.0625	–125.5625	154.1250	200.2150	–29.285 ÷ 4 = –7.32125
Correction		7.3213	7.3213	7.3213	7.3213	
		–250.741	–118.241	161.446	207.536	0
Round to:		–251	–118	161	208	0

To correct for seasonality:

Year	Quarter			$A - S$
1996	1	7554	–251	7805
	2	7817	–118	7935
	3	8101	161	7940
	4	8330	208	8122
1997	1	7994	–251	8245
	2	8338	–118	8456
	3	8795	161	8634
	4	8967	208	8759
1998	1	8559	–251	8810
	2	8727	–118	8845
	3	9111	161	8950
	4	9400	208	9192
1999	1	9041	–251	9292
	2	9248	–118	9366
	3	9731	161	9570
	4	9742	208	9534
2000	1	9616	–251	9867
	2	9891	–118	10009
	3	10857	161	10696
	4	9286	208	9078

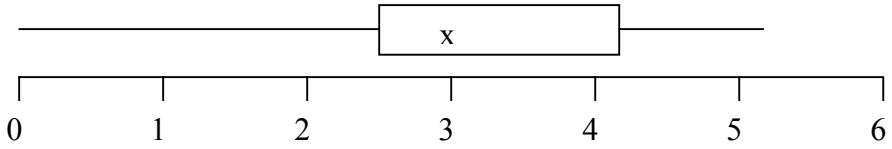
(ii) There is a strong seasonal pattern, in addition to the trend; Q3 is the main holiday period and Q4 includes pre-Christmas travel. The final figure, 2000 Q4, is an outlier on the evidence of remaining data, so some attempt to find an explanation is needed.

(iii) With such a rapid increase in the actual figures, it may be that the seasonal use has increased proportionately so that a multiplicative model would be better. However, the $A - T$ figures above do not really suggest this is the case. A multiplicative model may be analysed following a log transformation to make it linear (or in other ways also).

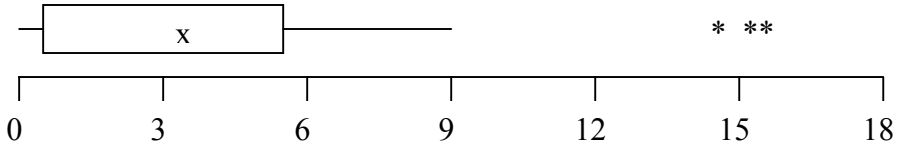
Higher Certificate, Paper III, 2002. Question 5

(i) Boxplots require median, quartiles, minimum and maximum values. As each set of data has been arranged in increasing order of size, it is easy to check whether any outliers have been included when making these calculations.

For Banks, the listing shows no outliers. Since $N = 36$, the median M is between the 18th and 19th observations, both of which are 3.0. Q_1 is between the 9th and 10th, which are 2.4 and 2.5. The program calculates 2.425; 2.45 is also acceptable. Q_3 is found in a similar way.

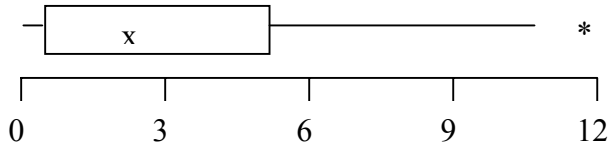


For E&EEq there are three obvious outliers. $N = 46$, so M is between the 23rd and 24th observations, at 3.25. Clearly the full set of data have been used in the program's calculation (not the $N = 43$ when outliers are omitted). Q_1, Q_3 are 0.3, 5.325, or near to these depending on the method of calculation.



The * points are outliers, and the upper whisker does not include them. [Since we use median and quartiles, not mean and standard deviation, the difference between values with and without outliers would not be great.]

For SS, there may be one outlier at 11.9 (although there is also considerable skewness at the upper end). $M = 2.2$, $Q_1 = 0.325$, $Q_3 = 4.775$.



*[**Note.** The plots drawn above might not appear **exactly** correct, due to screen and/or printer resolution.]*

Continued on next page

(ii) The column of numbers 001122… is the "stem", which is the number before the decimal point. To the right of this, listed in increasing order, are the "leaves", which are the decimal parts; those with decimal parts 0 – 4 are listed on one row, and 5 – 9 on the next row below; for example, $\begin{matrix} 0\ 0 \\ 0\ 789 \end{matrix}$ shows that the first four data in 'Banks' were 0.0 0.7 0.8 0.9.

Frequencies are cumulated, row by row, from <u>each</u> end, so that they meet in the middle, where the bracketed number, e.g. (7) for Banks, shows the actual frequency in the interval containing the median.

The dotplots show where each observation is located on the scale of measurement, with one dot for each item. (On the scales used here, spacing forces some adjacent numbers to come together.)

(iii) <u>Banks</u>: There are 4 very low figures, 1% or less; apart from these, there is a reasonably symmetrical pattern with 3% as its approximate centre. The spread of this set of data is not great; there are no upper outliers and the range of the 32 items excluding the 4 low ones is from 2.1 to 5.1.

<u>E&EEq</u>: The DESCRIBE program results are rather distorted by the three very large observations, which are clearly outliers, and the general skewness of the whole pattern. There is a substantial number of zeros. An "exponential decay" pattern (an exponential distribution) might explain all but the last three.

<u>SS</u>: This is rather similar to E&EEq, having several zeros and an exponential pattern. However, this pattern gradually tails off and does <u>not</u> have outliers far above their neighbours – it is probably wise to treat 11.9 as an outlier, although it would fit an exponential distribution.

<u>General</u>: Stem and leaf diagrams seem to give the interpretation more easily than dotplots (at least with the program used) but the numerical descriptive summaries seem less satisfactory than either of these graphical methods. The question of a sub-group of zeros arises in two of the data sets.

Higher Certificate, Paper III, 2002. Question 6

Note that there about 5600 members in Grade I, and 1400 in Grade II; also about 5250 in areas ABC and 1750 in the rest of the world.

Simple Random Sampling from the alphabetical list of members would be easy to organise; questionnaires could be distributed separately or perhaps by including them in the appropriate copies of the next issue of the journal (or in any other regular publication such as a newsletter). It may not be a very good method because the Grade II members are a small proportion, as are "rest of the world" ones. These groups could be in danger of not being sampled very well.

Stratified Random Sampling would be less easy but far more satisfactory because not only these smaller groups but also the A/B/C groups could be examined satisfactorily according to likely variability, cost, proportion satisfied, as well as having appropriate numbers from each group. Lists subdivided more than just by grade would be useful, and modern data storage methods should make identifiers for subgroups easy to provide.

Quota Sampling is totally infeasible. It would be very desirable to split into several groups as suggested above, and if it were possible quota sampling would produce the required sample numbers. As it is not possible, reminders to non-respondents would be the only way of achieving reasonable sample sizes in subgroups.

Cluster Sampling is not feasible because there is no obvious way of splitting into clusters, nor could enough of them be produced to make sampling from them a reasonable process. There do not seem to be any theoretical grounds for wanting to sample in clusters either.

Systematic Sampling from the original alphabetic list would be very easy, and probably just as satisfactory as simple random sampling. However, there is a distinct risk that some surnames would be especially associated with some areas, so that stratification would be better. If the sample method is going to involve producing lists in different groups to sample from, systematic sampling could be used instead of random choice because it may be quicker.

Possible Groups would be Grades I, II, each split into A, B, C, "rest". A 'good' method must compare Grades satisfactorily. This seems to be the most important requirement in the specification.

Higher Certificate, Paper III, 2002. Question 7

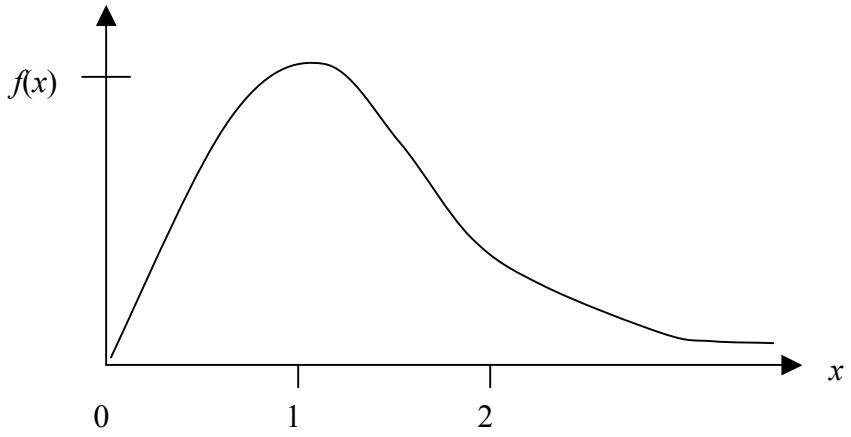
$f(x) = \lambda^2 x e^{-\lambda x}$, $x \geq 0$. Mean is $2/\lambda$. Cdf is $F(x) = 1 - (1+\lambda x)e^{-\lambda x}$, $x \geq 0$.

(i) $\dfrac{df(x)}{dx} = \lambda^2 \left\{ e^{-\lambda x} - \lambda x e^{-\lambda x} \right\}$ which is 0 for $(1 - \lambda x) = 0$, i.e. $x = \dfrac{1}{\lambda}$. This is the mode. Check that $\dfrac{d^2 f(x)}{dx^2}$ is negative here, so that we do indeed have a maximum:

We have $\dfrac{df(x)}{dx} = \lambda^2 e^{-\lambda x}(1 - \lambda x)$, so $\dfrac{d^2 f(x)}{dx^2} = \lambda^2 e^{-\lambda x}(-\lambda) + \lambda^2 (1 - \lambda x)\left(-\lambda e^{-\lambda x}\right)$

and at $x = \dfrac{1}{\lambda}$ the second term is 0 and the first is < 0, so $\dfrac{d^2 f}{dx^2} < 0$.

(ii) When $\lambda = 1$, $f(x) = xe^{-x}$ and the mode is at $x = 1$. $f(1) = 0.3679$.



***[Note.** The diagram is not drawn to scale.]*

(iii) In general, the mode is at $1/\lambda$ and the mean is $2/\lambda$. The distribution is strongly skew to the right. It is not very likely that this distribution would work well in a supermarket that channels customers with only few ("less than 10", say) purchases through special cash outlets. But if there are none of these and everyone must go through the same channels, then there will be some very small values of x to combine with the higher ones, and with a few very high service times. This distribution could work if applied to pooled data from all the outlets together.

Continued on next page

(iv) $L(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} \lambda^2 x_i e^{-\lambda x_i} = \lambda^{2n} e^{-\lambda \sum x_i} \left(\prod x_i \right)$

and $l = \ln(L) = 2n \ln \lambda - \lambda \sum x_i + \ln\left(\prod x_i \right)$.

So we have $\frac{dl}{d\lambda} = \frac{2n}{\lambda} - \sum x_i$; $\frac{dl}{d\lambda} = 0$ gives $\frac{2n}{\hat{\lambda}} = \sum x_i$ or $\hat{\lambda} = \frac{2}{\bar{x}}$.

$\frac{d^2 l}{d\lambda^2} = -\frac{2n}{\lambda^2} < 0$, so this is a maximum.

Setting $\bar{x}$ equal to $E[X] = \frac{2}{\lambda}$, we find that the moments estimator of λ is $\frac{2}{\bar{x}}$. In this case, both estimators are the same.

(v) $F(x) = 1 - (1 + x)e^{-x}$, and n = 200 service times are sampled.

$F(1.0) = 1 - 2e^{-1} = 0.2642$; and $F(1.5) = 1 - 2.5e^{-1.5} = 0.4422$.

So $P(1.0 < X \leq 1.5) = 0.4422 - 0.2642 = 0.1780$, and the corresponding expected frequency is 35.59.

Using the fact that 200 observations were made, the final value for $x > 5.0$ is 8.09.

The usual chi-squared test here will have 10 degrees of freedom; no parameters had to be estimated, and all 11 intervals can be used since only one has an expected value that is just below 5. The test statistic is

$$\frac{(21-18.04)^2}{18.04} + \frac{(30-34.81)^2}{34.81} + \frac{(36-35.59)^2}{35.59} + \frac{(29-30.36)^2}{30.36} + \frac{(27-23.74)^2}{23.74}$$
$$+ \frac{(19-17.63)^2}{17.63} + \frac{(17-12.65)^2}{12.65} + \frac{(4-8.86)^2}{8.86} + \frac{(8-6.10)^2}{6.10} + \frac{(2-4.13)^2}{4.13} + \frac{(7-8.09)^2}{8.09}$$

= 7.769.

This value is nowhere near to being significant when compared with χ^2_{10}, so there is no evidence to reject the given null hypothesis.

Higher Certificate, Paper III, 2002. Question 8

(i) Completely randomised design is analysed according to the linear model

$$y_{ij} = m + t_i + e_{ij}, \quad \text{where } i = 1 \text{ to } 4,\ j = 1 \text{ to } 10,$$

y_{ij} is the number of weeds on the jth plot that received treatment i, t_i is the effect (departure from overall mean m) due to treatment i, and e_{ij} is a random (natural variation) term, Normally distributed with mean 0 and variance σ^2 which is constant for all observations.

The variances in the four treatments do not appear constant in the untransformed data.

We assume that the model is additive (a sum of terms) but cannot check this without computing the values of the residuals.

(ii) The transformation $\exp\left(\frac{y}{100}\right)$ also gives very unequal variances. The range of variances in $\sqrt{y}$ is largest/smallest $\approx$ 8.7 whereas for $\log_{10}(y)$ it is ≈ 4.0; thus we should choose $\log_{10}(y)$. However, a ratio 4:1 among variances is still rather high, though not unusual with small samples of data. A better transformation could probably be found, provided it made physical sense.

(iii) The herbicide totals using $\log_{10}(y)$ are 19.857, 18.943, 21.260, 22.502, each based on 10 observations; these add to 82.562. Herbicides SS is therefore

$$\frac{1}{10}\left(19.857^2 + \ldots + 22.502^2\right) - \frac{1}{40}\left(82.562^2\right) = 171.1465302 - 170.4120961 = 0.7344341.$$

Analysis of Variance of $\log_{10}(y)$:

Source	df	Sum of Squares	Mean Square	*F* value
Herbicides	3	0.73443	0.2448	19.45 (very highly sig)
Residual	36	0.45300	0.01258	
Total	39	1.18743		

There is strong evidence of the presence of differences among the herbicide means, since the value 19.45 is very highly significant when referred to $F_{3,36}$.

(iv) A residual is the difference between an observed y_{ij} and its fitted value using the linear model. In a completely randomised design, fitted values for each treatment are the treatment mean for that treatment. If residuals are plotted against fitted values, for each observation, the resulting pattern should show a set of values randomly scattered about 0, with concentration near 0 and no outliers so that the Normality assumption is acceptable. Variability should show no pattern depending on the size of y_{ij}, or which treatment it received.