THE ROYAL STATISTICAL SOCIETY

2003 EXAMINATIONS – SOLUTIONS

HIGHER CERTIFICATE

PAPER II – STATISTICAL METHODS

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 II, 2003. Question 1

(i) The chi-squared test will examine the null hypothesis that there is no relation between "Commercial" and "Purchase", against the alternative hypothesis that there is a relation.

Observed frequencies and (in brackets) expected frequencies on the null hypothesis:

		Commercial		
		A	*B*	*Total*
Purchase	*No*	70 (75)	80 (75)	150
	Yes	30 (25)	20 (25)	50
	Total	100	100	200

$$\text{Test statistic} = \frac{(70-75)^2}{75}+\frac{(80-75)^2}{75}+\frac{(30-25)^2}{25}+\frac{(20-25)^2}{25} = 2.667$$

[or 2.16 if calculated with Yates' correction so that "$(O-E)^2$" becomes $(4.5)^2$].

Refer to χ_1^2: not significant. There is no evidence of a relation between "Commercial" and "Purchase".

Hence a decision could be made on non-statistical grounds, such as cost of the commercial or the potential size of the audience.

(ii) McNemar's test for paired data tests similar hypotheses on association or otherwise of the two classifications, in this case which advertising medium each manufacturer uses. It does not use either "No–No" or "Yes–Yes" manufacturers.

Test statistic $= \dfrac{(5-15)^2}{5+15} = \dfrac{100}{20} = 5.00$, refer to χ_1^2, significant at 5%. There is evidence against a null hypothesis of no preference of advertising medium.

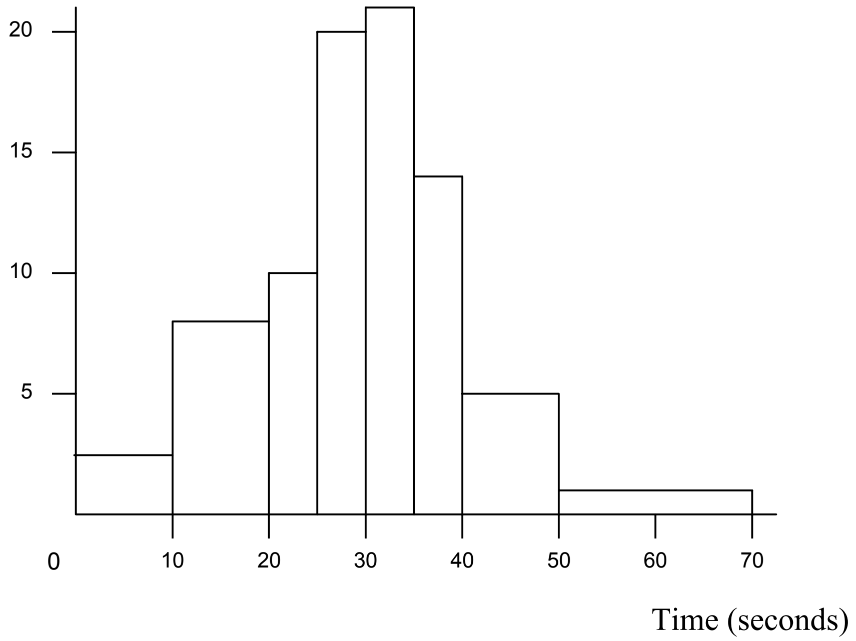
(iii) In part (i), two different random samples of responses are obtained, and the problem is that of comparing the proportions giving a particular response in the two samples. This is often the case in a chi-squared test, for example in opinion surveys where the two samples are drawn from males and females, or "young" and "old" age-groups, in otherwise similar populations. It also applies in medical trials where different groups of patients (e.g. smokers and non-smokers) are classified as having or not having a particular disease.

However, in part (ii) there are not two independent samples, and this may occur more often in medical trials; for instance, suppose two drugs are used to treat a chronic illness, each for a short period of time, on the same patients (or at the least on patients who have been closely paired for age, sex and general medical condition). No information is gained from patients (or pairs) where both drugs worked, or both failed. The McNemar test examines whether, in cases where only one worked, drug A was more successful than drug B or not. It compares the proportions of preferences in this part of the data only. The example in part (ii) uses the same manufacturers, so a McNemar test is valid.

Higher Certificate, Paper II, 2003. Question 2

Time (t)	Frequency (f)	Midpoint of time interval (x)	Cumulative frequency (F)	fx	fx^2
0 – 10	5	5	5	25	125
10 – 20	16	15	21	240	3600
20 – 25	10	22.5	31	225	5062.5
25 – 30	20	27.5	51	550	13125
30 – 35	21	32.5	72	682.5	22181.25
35 – 40	14	37.5	86	525	19687.5
40 – 50	10	45	96	450	20250
50 – 70	4	60	100	240	14400
	100			2937.5	100431.25

(i) Frequency density (per 5 seconds)



(ii) Mean = $\dfrac{2937.5}{100}$ = 29.38 seconds. Median = $30 - \left(\dfrac{1}{20}\times 5\right)$ = 29.75 seconds.

The distribution is roughly symmetrical, perhaps a little skewed to the left.

Variance $s^2 = \dfrac{1}{99}\left(100431.25 - \dfrac{(2937.5)^2}{100}\right) = 142.8504$, so $s = 11.95$.

(iii) Assuming Normality of the distribution of times, and using the large-sample formula, 95% limits are

$29.38 \pm 1.96\times\dfrac{11.95}{\sqrt{100}}$, i.e. 29.38 ± 2.34 or (27.04, 31.72) seconds.

Higher Certificate, Paper II, 2003. Question 3

Null hypotheses to be tested are $\sigma = 0.05$ and $\mu = 10.5$.

Summary statistics for the two sets of apparatus are:

A: $\bar{x} = 10.48$, $s^2 = 0.008898$ ($s = 0.09433$)

B: $\bar{x} = 10.31$, $s^2 = 0.003876$ ($s = 0.06228$)

$n = 8$ in both cases.

(i) Alternative hypothesis is $\sigma^2 > (0.05)^2 = 0.0025$. Test statistic is $\dfrac{(n-1)s^2}{\sigma^2}$, refer to χ^2_{n-1}, i.e. χ^2_7 here: upper 5% point is 14.07, upper 1% point is 18.48.

For A, we get $\dfrac{7 \times 0.008898}{0.0025} = 24.91$, highly significant. For B, we get $\dfrac{7 \times 0.003876}{0.0025} = 10.85$, not significant.

The null hypothesis is rejected for A, but cannot be rejected for B. There is evidence that A is more variable than standard, but not that B is more variable.

(ii) Alternative hypothesis is $\mu \neq 10.5$. Test statistic is $\dfrac{\bar{x} - 10.5}{s/\sqrt{n}}$, refer to t_{n-1}, i.e. t_7 here.

For A, we get $\dfrac{-0.02\sqrt{8}}{0.09433} = -0.60$, not significant. For B, we get $\dfrac{-0.19\sqrt{8}}{0.06228} = -8.63$, extremely highly significant.

There is very strong evidence that B is biased (downwards) but none that A is biased.

(iii) Probably B only needs a scale of measurement adjusted, if the complete process is automated; more seriously there may be a fault in the way the potassium content is measured. For A, there is too much variation, though the mean value is acceptable, and this is likely to need adjustment to that part of the process which can give rise to variability. In each case a laboratory technician should be called in.

Higher Certificate, Paper II, 2003. Question 4

(i) A sign test is appropriate, the null hypothesis being that A and V are equally likely to be the greater. Hence the number of As is binomial with parameters 10 and ½, as is the number of Vs, if the null hypothesis is true. The alternative hypothesis claims that V is greater. A one-sided test is therefore needed.

The observed results are $n_A = 3$, $n_B = 7$.

If the null hypothesis is true, we have

$$P(n_V \geq 7) = \left\{ \binom{10}{7} + \binom{10}{8} + \binom{10}{9} + 1 \right\} \frac{1}{2^{10}} = \frac{120+45+10+1}{2^{10}} = 0.172 \ .$$

There is not enough evidence to reject the null hypothesis.

The sign test is not very powerful. The sample size here (8) is not really large enough for its effective use.

(ii) (a) The data provide information which the sign test would not use, namely that measuring the change on a numerical scale. As well as the sign of the change we should use the size. A Wilcoxon signed-rank test is suitable. Difference $A - B$ are as follows, and the absolute values of the differences are ranked in size order, with tied values given their average ranking.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	7	–2	6	4	24	15	–5	1	3	5	–7	–1
Rank	9½	3	8	5	12	11	6½	1½	4	6½	9½	1½

The null hypothesis is that aural memory scores are not altered by coaching, the alternative hypothesis is that coaching leads to improvement. A one-sided test is required.

The sum of ranks of negative values is $T_- = 20½$ and of positive values is $T_+ = 57½$. We use $T = \min(T_-, T_+) = 20½$. Tables for $n = 12$ give 17 as the critical value for a one-sided 5% test, so there is not enough evidence to reject the null hypothesis.

(b) A paired-samples t test using the differences would be appropriate if the distribution of differences appeared to be approximately Normal. The two large values for (5) and (6) make this unlikely, both being on the same side (+). It would be unwise to use the t test in this case.

Higher Certificate, Paper II, 2003. Question 5

(i) When a large sample $\{x_i\}$ of data is available from any distribution (continuous or discrete) whose mean is μ and (finite) variance is σ^2, the total ΣX_i and the mean of the sample, $\overline{X}$, are both approximately Normally distributed with parameters ($n\mu$, $n\sigma^2$) for the total and (μ, σ^2/n) for the mean. The sample size n is required to be "large", but the implication of this for data collection depends on the shape of the distribution; if it is not too unsymmetrical n can be quite small, but for a highly skew distribution n needs to be very large.

For example, data from agricultural plots, consisting of a large number of individual plants, can usually be treated as approximately Normal, and so can data from large-scale surveys. This allows statistical inference based on the theory for the Normal distribution to be used. It also allows maximum likelihood estimators based on large samples to be treated as Normal, for example in constructing confidence intervals.

(ii) For the new drug, $n_1 = 144$, $\bar{x}_1 = 50.6$, $s_1^2 = 10.3$;
For the placebo, $n_2 = 144$, $\bar{x}_2 = 35.4$, $s_2^2 = 14.7$.

The variance of $\left(\overline{X}_1 - \overline{X}_2\right)$ is $\left(\sigma_1^2/n_1\right)+\left(\sigma_2^2/n_2\right)$, and with large samples (as these are for this type of measurement) we simply use s_1^2 and s_2^2 for σ_1^2 and σ_2^2, and use the Normal approximation to obtain the interval

$$\left(\bar{x}_1 - \bar{x}_2\right) \pm 1.96\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} .$$

Thus we get

$$\left(50.6 - 35.4\right) \pm 1.96\sqrt{\frac{10.3+14.7}{144}} ,$$

$$\text{i.e. } 15.2 \pm 1.96\sqrt{\frac{25}{144}}, \quad \text{i.e. } 15.2 \pm 1.96\times\left(\frac{5}{12}\right),$$

which is 15.2 ± 0.82 or (14.38, 16.02).

The new drug appears to give (with 95% confidence) between 14.38 and 16.02 extra hours of sleep in the week. This is a substantial improvement, of at least 2 hours per day, and significant at a high level since the interval does not contain the value 0 (or go anywhere near it).

Continued on next page

(iii) Treatment A: $n_A = 250$, proportion improving $\hat{p}_A = 205/250 = 0.82$.
Treatment B: $n_B = 250$, proportion improving $\hat{p}_B = 180/250 = 0.72$.

These samples are large enough to use a Normal approximation to the distributions of $\hat{p}_A$ and $\hat{p}_B$, and $\text{Var}\left(\hat{p}_A - \hat{p}_B\right)$ can be estimated as

$$\frac{\hat{p}_A\left(1-\hat{p}_A\right)}{n_A}+\frac{\hat{p}_B\left(1-\hat{p}_B\right)}{n_B} = \frac{0.82\times0.18}{250}+\frac{0.72\times0.28}{250} = \frac{0.1476+0.2016}{250}$$

$= 0.0013968$, so we have $\text{SE}\left(\hat{p}_A - \hat{p}_B\right) = 0.0374$.

Also, we have $\hat{p}_A - \hat{p}_B = 0.10$, and therefore 95% limits for $p_A - p_B$ are (approximately)

$0.10 \pm \left(1.96\times0.0374\right)$, i.e. 0.10 ± 0.073 or $(0.027, 0.173)$.

Treatment A gives better results than B, by an amount between 2.7% and 17.3% improvement (with 95% confidence). This is a significant improvement since the interval does not contain 0.

Higher Certificate, Paper II, 2003. Question 6

(i) $y_{ij} = m + t_i + e_{ij}$,

where y_{ij} is the measurement made on the jth unit receiving the ith treatment, m is the underlying population mean of all observations and $\{e_{ij}\}$ are independent, Normally distributed, residual (natural) variation terms with mean 0 and common variance σ^2.

This model separates the total variation among the $\{y_{ij}\}$ into a systematic component due to "treatments" t_i and a random component represented by e_{ij}. The number of replicates of treatment i is r_i [i.e. we have $j = 1, 2, \ldots, r_i$ for each i], and $\Sigma r_i = N$, the total number of experimental units.

(For the usual form of analysis, $\{e_{ij}\}$ are assumed to be Normal, although randomisation theory validates the inferences usually made from a one-way analysis.)

(ii) (a) $r_i = 5$ for $i = A, B, C, D$. $N = 20$. $\Sigma\Sigma y_{ij}^2 = 13666$.

Treatment (fertiliser) totals are A 112, B 161, C 119, D 124.

Grand total $G = 516$.

Corrected total SS $= \Sigma\Sigma y_{ij}^2 - (G^2/N) = 13666 - \{(516)^2/20\}$

$= 13666 - 13312.8 = 353.2$.

SS for fertilisers

$$= \sum \frac{T_i^2}{r_i} - \frac{G^2}{N} = \frac{112^2 + 161^2 + 119^2 + 124^2}{5} - \frac{G^2}{N}$$

$$= \frac{68002}{5} - \frac{G^2}{N} = 13600.4 - 13312.8 = 287.6.$$

Analysis of variance

ITEM	df	Sum of Squares	Mean Square	F ratio
Fertilisers	3	287.6	95.87	23.38
Residual	16	65.6	4.10	
Total	19	353.2		

The F ratio is referred to $F_{3,16}$ and is very highly significant, leading us to reject a null hypothesis that there are no differences between fertiliser mean yields.

Continued on next page

Fertiliser means are A 22.4, B 32.2, C 23.8, D 24.8 (kg).

σ^2 ("residual variation" or "experimental error") is estimated as 4.10.

Significant differences can be claimed between any pair of means differing by at least $t_{(16)}\sqrt{\dfrac{2\hat{\sigma}^2}{5}} = t_{(16)}\sqrt{1.64}$ or $1.281t_{(16)}$ where $t_{(16)}$ is the two-tailed 5% point of t_{16}, i.e. 2.120. Thus $1.281t_{(16)} = 2.71$.

Clearly B is different from all the others and there are no differences between A, C, D. Assuming that all the four fertilisers were applied in the same way, at the same time, it is reasonable to claim that B is best.

The farmer simply needs to be told that statistical analysis very strongly suggests that the four fertilisers did not all give similar results, and that after allowing for the natural variations among the crop we can say that B is clearly better than A, C, D.

(b) Five randomised blocks, the columns in the diagram on the question paper, should be used. This will remove an "east–west" trend. In each column, one replicate of each treatment (fertiliser) should be set out in random order (different randomisations being used for each block).

Higher Certificate, Paper II, 2003. Question 7

Suitable diagrams include the following:

> pie charts, for chosen years between 1990 and 2001, showing the distribution of spending between different categories of the household expenditure;
>
> time series graphs for individual categories, expressing expenditure either as an absolute figure or a percentage of the total;
>
> bar charts, similar in purpose to pie charts, either as % bar charts or totals to show overall expenditure as well as components.

Because prices are given in terms of 2000/2001 levels, it may be less easy to track the effects of price changes on consumption of individual items, or to see what may have altered in the expenditure of items within each category.

Some points suggested by the raw data are:

> total was steady early on, then began to increase, more quickly at the end;
>
> housing showed a fall, then a rise, sharp at the end;
>
> fuel and power fell, particularly from 1998 on;
>
> household and personal goods and services increased in absolute value, as did travel and leisure;

These latter four categories could certainly be explored as percentages of total, as well as absolute values.

A newspaper article might emphasise the largest and smallest areas of spending, in which areas spending increased, decreased or remained constant, and any apparent relations in behaviour of the various categories (e.g. the four noted above).

Higher Certificate, Paper II, 2003. Question 8

(i) The F distribution with m and n degrees of freedom is the ratio of two independent chi-squared distributions divided by their numbers of degrees of freedom:

$$F_{m,n} = \frac{\dfrac{\chi_m^2}{m}}{\dfrac{\chi_n^2}{n}} .$$

Thus if estimates of variance s_1^2 and s_2^2 are obtained from each of two independent samples of Normally distributed data, (m+1) and (n+1) items in the samples respectively, the null hypothesis that the underlying population variances σ_1^2 and σ_2^2 are equal can be tested by referring s_1^2/s_2^2 to $F_{m,n}$. A confidence interval for σ_1^2/σ_2^2 can also be found.

In analysis of variance, sums of squares will have χ^2 distributions if Normality is assumed. A suitable null hypothesis is set up, such as that a regression coefficient is zero. The appropriate sum of squares is then compared with the residual sum of squares, using the "mean squares" so as to deal with the numbers of degrees of freedom, using an F distribution. If the null hypothesis is true, both the mean squares will merely estimate experimental error so their ratio has expected value 1, but if it is false the mean square corresponding to the regression coefficient is expected to be larger. One-way and two-way (and higher-way) analysis of variance for experimental designs uses the same background theory.

(ii) The null hypothesis is "$\sigma_A^2 = \sigma_B^2$", the same variability for both machines, and the alternative hypothesis is "$\sigma_A^2 \neq \sigma_B^2$".

Summary statistics for the two machines are:

$$A: \quad n_1 = 16, \;\; \bar{x}_1 = 1000.125, \;\; s_1^2 = 10.9167 \;\; (s = 3.304);$$
$$B: \quad n_2 = 20, \;\; \bar{x}_2 = 1000.050, \;\; s_2^2 = 2.9974 \;\; (s = 1.731).$$

s_1^2/s_2^2 = 3.64, refer to $F_{15,19}$. This is significant at the 5% level so the null hypothesis can be rejected and we can decide to use B because the evidence is that it is less variable than A.

A report should mention that both machines dispense, on average, just over 1000 ml. However, the data from the trial period show that A was significantly more variable than B (in fact, inspection of the data shows two rather outlying values in A, at 993 and 1007), so B is the one to buy.