THE ROYAL STATISTICAL SOCIETY

2004 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, 2004. Question 1

Part (i)

All four tests are for comparing two samples of data.

(a) If two samples from Normal distributions are available, and it can be assumed that their population variances are the same but this value is not known, we can compare means using a *t* test. The null hypothesis under test is $\mu_1 = \mu_2$, where these are the population means.

Examples are commonly of two independent samples from the same basic population but which have been "treated" in different ways – such as plants in an agricultural experiment with different fertilisers, or people of similar IQ in an educational trial with different teaching methods.

(b) This introduces "blocking" of the experimental units in pairs to remove some possible systematic variation in experimental material.

For example, an experiment to compare two "treatments" for plants in a glasshouse might have adjacent pairs of plants, the two in each pair thus encountering ambient conditions as nearly alike as possible, one being "treated" in one way and the other in the other way. Differences *within* the pairs can then reasonably be ascribed to differences between the treatments; possible variations in ambient conditions in different parts of the glasshouse would not affect the within-pairs comparisons. (Of course, the two plants within each pair should be as nearly alike as possible in the first place.) The population of differences between responses within the pairs has to be Normally distributed and the null hypothesis is that the mean of this population is zero (which is equivalent to the means of the two separate populations of responses being equal).

(c) This test also compares two "treatments" but using rankings.

As an example, suppose each member of a group of people, chosen to be as similar as possible, is asked to carry out a computer task under one of two different sets of background conditions, and their accuracies are ranked 1, 2, 3, … as a single ordering. The null hypothesis is that the two underlying populations are the same, the alternative being that they differ in location. The null hypothesis is equivalent to the single ordering of ranks being in random order as far as the "treatments" (conditions) are concerned. [In contrast, if "treatment" *A* was better than "treatment" *B*, we would anticipate that *A* would tend to have high ranks (if "high" means "better") so that, with respect to the "treatments", the single ordering would have mainly *B*s at the start and *A*s at the end.] The test is based on the sum of the ranks for each "treatment". No background distributional assumptions are required.

(d) This, like (b) above, is a paired test. It is carried out for the same general reason of removing possible systematic variation in experimental material.

As an example, suppose that the concentration of a chemical substance in the blood is measured on the same people before and after receiving a drug treatment. There may be wide variations from person to person, but each before-and-after comparison for the *same* person should give a good indication of the effect, if any, of the drug. A suitable null hypothesis here is that there is no change in concentration, and as in (c) no background distributional assumptions are required. The test is based on ranking the before-and-after differences (absolute values) and calculating the rank sums for positive and negative differences.

Continued on next page

Part (ii)

With such small sets of data, it is not clear whether we should assume Normality. A dot-plot (or Normal probability plot if available) might shed some light. The choice is between (i)(a) above if Normality is assumed and (i)(c) above if not. It turns out in this case (see below) that both tests give similar inferences.

Under (i)(a)

We must first compare variances, to check the assumption of equal population variances. We have that for A, $\bar{x}_A = 59.00$, $s_A^2 = 40.3333$; and for B, $\bar{x}_B = 67.57$, $s_B^2 = 27.9524$. $\therefore s_A^2 / s_B^2 = 1.44$, refer to $F_{6,6}$ – not significant, so it is reasonable to assume that the population variances are the same.

Thus we may calculate the pooled s^2 which is 34.1429.

Test statistic for testing $\mu_A = \mu_B$ against $\mu_A \neq \mu_B$ is

$$\frac{\bar{x}_A - \bar{x}_B(-0)}{s\sqrt{\frac{1}{7}+\frac{1}{7}}} = -\frac{8.57}{3.12} = -2.74,$$

which is referred to t_{12}. This is significant at the 5% level (double-tailed 5% point is 2.179, double-tailed 1% point is 3.055), so there is evidence to reject the null hypothesis – it seems that the population means are not the same (and that the mean for A is lower than that for B).

Under (i)(c)

The joint ranking is as follows.

Score	50	52	58	59	60	61	62	64	67	68	69	71	72	73
Rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Test	*A*	*A*	*A*	*B*	*A*	*A*	*B*	*A*	*B*	*A*	*B*	*B*	*B*	*B*

$n_1 = 7$, $n_2 = 7$. Total of ranks for $A = 35$; total of ranks for $B = 70$.

We test the null hypothesis that the two underlying populations are the same against the alternative that they differ in location.

Calculating the Mann-Whitney statistic via the ranks (note: it can also be calculated directly, or the Wilcoxon rank-sum form could be used),

$$U_1 = n_1 n_2 + \tfrac{1}{2} n_1 (n_1 + 1) - R_A = 49 + 28 - 35 = 42.$$

$$U_2 = n_1 n_2 + \tfrac{1}{2} n_2 (n_2 + 1) - R_B = 49 + 28 - 70 = 7.$$

[Equivalently, these can be calculated as $35 - \tfrac{1}{2} n_1 (n_1 + 1) = 35 - 28 = 7$ and $70 - \tfrac{1}{2} n_2 (n_2 + 1) = 70 - 28 = 42$.]

So $U_{min} = 7$. From tables, the critical value for a U test with $n_1 = n_2 = 7$ at the 5% two-tailed level is 9. As $7 < 9$, there is evidence to reject the null hypothesis. Noting that it is A that has the lower ranks, it seems that the location for the A population is lower than that for the B population.

Higher Certificate, Paper III, 2004. Question 2

(i) Useful questions would be (1) were the files read using the same hardware under the same conditions, (2) were the data given to a fixed number of digits and to the same accuracy, (3) was there any "competition" by other network users (which could slow down the reading time).

(ii) $y_{ij} = \mu + \tau_i + \varepsilon_{ij}$, where y_{ij} is the jth reading of format i (here i goes from 1 to 3, j from 1 to 8 for each i), μ is the overall population general mean, τ_i the population mean effect due to being in format i. The Normally distributed residual (error) terms ε_{ij} all have variance σ^2 and are uncorrelated (independent). Random sampling should lead to independence. Normality is not easy to check in small samples; dot-plots or box and whisker plots could be used (or Normal probability plots if available). Constant variance is also hard to check in small samples as tests are not very sensitive; dot-plots sometimes give useful information.

The format totals and means are:			
	Standard	17.19,	2.149
	First	16.16,	2.020
	Second	18.52,	2.315

The grand total is 51.87. $\Sigma\Sigma y_{ij}^2 = 113.3981$.

"Correction factor" is $\dfrac{51.87^2}{24} = 112.1040$.

Therefore total SS = 113.3981 – 112.1040 = 1.2941.

SS for formats = $\dfrac{17.19^2}{8} + \dfrac{16.16^2}{8} + \dfrac{18.52^2}{8} - 112.1040 \;=\; 0.3500$.

The residual SS is obtained by subtraction.

SOURCE	DF	SS	MS	*F* value
Formats	2	0.3500	0.1750	3.89 Compare $F_{2,21}$
Residual	21	0.9441	0.0450	$= \hat{\sigma}^2$
TOTAL	23	1.2941		

The upper 5% point of $F_{2,21}$ is 3.47, 1% point 5.78; the treatments effect is significant at the 5% level. There is some evidence that the null hypothesis, that mean times for the formats are all equal, should be rejected.

(iii) Null hypothesis: $\mu_{\text{FIRST}} = \mu_{\text{SECOND}}$. Alternative hypothesis: $\mu_{\text{FIRST}} \neq \mu_{\text{SECOND}}$.

We have $\bar{y}_F - \bar{y}_S = -0.295$, and the estimated standard deviation for this comparison is $\sqrt{2\hat{\sigma}^2/8} = 0.106$. So the t_{21} test statistic is –0.295/0.106 = –2.78, which is approaching significance at the 1% level (note: this comparison is likely to be the main reason for the significance of the overall *F* test in the analysis of variance above). A 95% confidence interval is given by –0.295 ± (2.08 × 0.106), i.e. (–0.515, –0.075) [i.e. the interval gives FIRST between 0.515 and 0.075 *less* than SECOND].

Higher Certificate, Paper III, 2004. Question 3

(i) By using the same subjects on both occasions, experiment 1 should give more precise results than experiment 2; subject-to-subject variation has been designed out. This assumes, of course, that any effect of the drug would indeed have worn off within the week.

(ii) $n = 10$. Differences d_i (drug – placebo) are 4, 3, 6, –1, 7, 0, –5, 8, 5, 5. So we have $\bar{d} = 3.2$, $s_d^2 = 16.40$. The required 95% confidence interval is given by $3.2 \pm \left(2.262 \times \sqrt{16.40/10}\right)$ where 2.262 is the double-tailed 5% point of t_9, i.e. the interval is (0.30, 6.10). We must assume that the differences are Normally distributed.

(iii) Let x refer to the drug and y to the placebo. We have $n_x = n_y = 10$. Sample means and variances are $\bar{x} = 197.1$, $s_x^2 = 816.322$ and $\bar{y} = 187.3$, $s_y^2 = 770.233$. We must assume that the two samples are from Normal distributions with the same variance.

The pooled estimate of this common variance is 793.278. The required 95% confidence interval is given by $(197.1 - 187.3) \pm \left(2.101 \times \sqrt{793.278\left(\frac{1}{10} + \frac{1}{10}\right)}\right)$ where 2.101 is the double-tailed 5% point of t_{18}, i.e. the interval is (–16.66, 36.26).

(iv) The interval in part (ii) does not contain zero; both its end-points are positive. This gives some evidence that there is an increase due to the drug. The interval in part (iii) is uninformative, being very wide and well spread in both directions about zero; subject-to-subject variation has not been designed out and thus inflates the estimate of variance.

Higher Certificate, Paper III, 2004. Question 4

(i)

Year	*Quarter*	*Sales*	*4-quarter totals*	*8-quarter totals*	*Moving average*
1997	1	31.54			
1997	2	22.33	104.46		
1997	3	20.29	105.27	209.73	26.216(25)
1997	4	30.30			
1998	1	32.35			

(ii) There is a sharp seasonal variation, "Sales – *MA*" being always substantially negative in quarters 2 and 3, always substantially positive in quarters 1 and 4. To estimate the pattern of seasonal variation, we need the average of the "Sales – *MA*" figures for each quarter.

	Q_1	Q_2	Q_3	Q_4		
1997			–5.93	3.73		
1998	5.42	–2.84	–6.38	3.90		
1999	6.24	–3.93	–5.88	3.95		
2000	5.15	–4.20	–4.78	3.88		
2001	5.77					
Seasonal totals	22.58	–10.97	–22.97	15.46		
Seasonal averages	5.645	–3.657	–5.743	3.865	Sum: 0.110	Correction: –0.028
Corrected seasonal averages	5.617	–3.685	–5.771	3.837	(–0.002)	

(iii) Using this case as an example, the visual pattern can show detail which is lost in the table of figures, such as in the year 2000 where the fluctuation was not so great as in other years, although the pattern was the same. We can also see that, while it rises overall, the *MA* trend shows a slight dip from late 1998 onwards before a sharper rise in 2000. We can visualise the trend from the table when it is fairly smooth like this, but not always so easily. Trend and seasonal variation are important properties to observe, and a clear method of doing so is invaluable.

(iv) We use observed sales minus estimated seasonal variation. For year 2000:

Q_1	Q_2	Q_3	Q_4
32.64 – 5.62 = 27.02	23.64 + 3.69 = 27.33	23.37 + 5.77 = 29.14	32.20 – 3.84 = 28.36

It would not be a good idea to use this method on the 2001 sales because the estimated seasonal variation might have changed if we had had enough data to make "Sales – *MA*" up to the end of the year.

Higher Certificate, Paper III, 2004. Question 5

(i) The total number of collisions rose steadily for the first three years and then fell in 1997. Total casualties rose in 1995, then dropped in 1996 and then rose again in 1997. Percentage changes from one year to the next were as follows.

	1994 to 1995	*1995 to 1996*	*1996 to 1997*
Total collisions	+6.9	+4.1	–2.8
Total casualties	+6.0	–8.0	+4.2

To consider seriousness of collisions, we might combine "fatal" and "major":-

	1994	*1995*	*1996*	*1997*
Number	30734	32323	30557	30849
% of total collisions	6.6	6.5	5.9	6.1

The actual numbers were very similar except for the increase in 1995, but in the last two years they were a smaller proportion of the total.

Similarly we might combine "killed" and "seriously injured", noting that the number of fatal casualties was least in 1997 whereas the number of seriously injured was least in 1994:-

	1994	*1995*	*1996*	*1997*
Number	46529	50036	48321	48993
% of total casualties	33.6	34.1	35.8	34.8

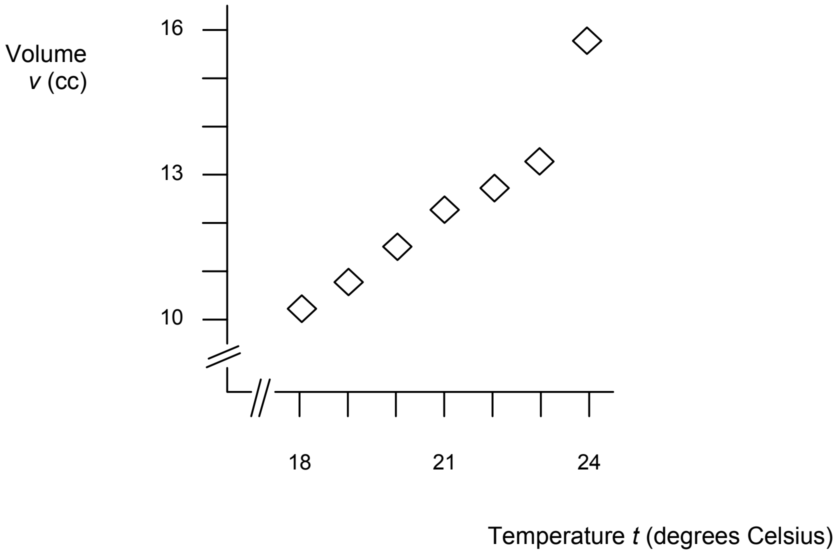
The actual numbers increased in 1995 but then steadied off; however, the percentages of the total increased for three years before falling slightly.

Useful diagrams would be component bar charts showing the annual totals of collisions and of casualties, with the components in each category (fatal etc, or killed etc) shown in different shading or colouring. Because annual changes are fairly small relative to total sizes, these diagrams will not show obvious or clear trends.

(ii) Examples of useful background information are

- occupancy of vehicles involved – e.g. driver only, few passengers, many passengers as in coaches
- type of vehicle – e.g. heavy lorry, car or other small vehicle (some collisions, e.g. between a lorry and a small car, seem more likely to cause fatalities)
- traffic density at the time
- numbers of registered vehicles in the year, indicating general level of road use / congestion
- mileage travelled by drivers involved
- ages of vehicles
- ages of drivers and length of experience
- roadworks or other local hazards.

(i)



Without the last point, there is an almost linear relation (perhaps a very slight flattening off?). The last point is considerably different (perhaps the measurement as recorded is an error for 13.90?).

(ii) The coefficients can be calculated using the usual linear regression formulae, but from the edited results each may be calculated directly as "T × SE Coef". Thus they are –5.51 and 0.852 respectively, so the fitted line is

"Volume = –5.51 + 0.852 × Temperature".

R^2 can be calculated as $S_{xy}{}^2/S_{xx}S_{yy}$ in the usual notation, i.e. here

$$\frac{\left(1842.03-\dfrac{147\times 86.58}{7}\right)^2}{\left(3115-\dfrac{147^2}{7}\right)\left(1092.9774-\dfrac{86.58^2}{7}\right)} = \frac{(23.85)^2}{28\times 22.1065} = 0.919 \text{ (or 91.9\%)}.$$

(iii) The regression omitting the last point gives a much better fit to the remaining points. This is reflected in the smaller residual mean square ($(0.05700)^2$ instead of $(0.5986)^2$) and hence the smaller standard errors of the coefficients, both of which give highly significant t statistics. R^2 is also greater for this regression. *If*, however, the last point is considered to be genuine and important, it is obviously not taken into account at all by the regression omitting it. The first regression does include it, but arguably a more complicated model than simple *linear* regression should be used anyway.

Continued on next page

(iv) As given by the second regression, volume increases by 0.657 cc for each 1 degree increase in temperature. The line, if projected back, would have $v = -1.68$ when $t = 0$ (this is of course absurd; obviously the linear regression would not hold that far outside the range of the available data).

R^2 is the proportion (usually given as a percentage, as in the edited results) of the total variation in the data that is explained by the regression relation.

"SE Coef" is the standard error of each regression coefficient. "T" is the value of each coefficient divided by its standard error (i.e. it is the value of the t test statistic for testing the hypothesis that the true value of the coefficient is zero). "P" is the probability of obtaining the calculated value of T or a more extreme value, on the null hypothesis that the true value of the coefficient is zero.

These values indicate the precision with which the line is fitted and allow the null hypothesis for each coefficient to be tested. Since $P < 0.05$ for each coefficient, we would reject each null hypothesis at the 5% level (indeed, the results are very highly significant, beyond the 5% level) and conclude that temperature does appear to (help to) explain volume.

In calculating P, a Normal distribution of the residual terms in the usual linear regression model is assumed.

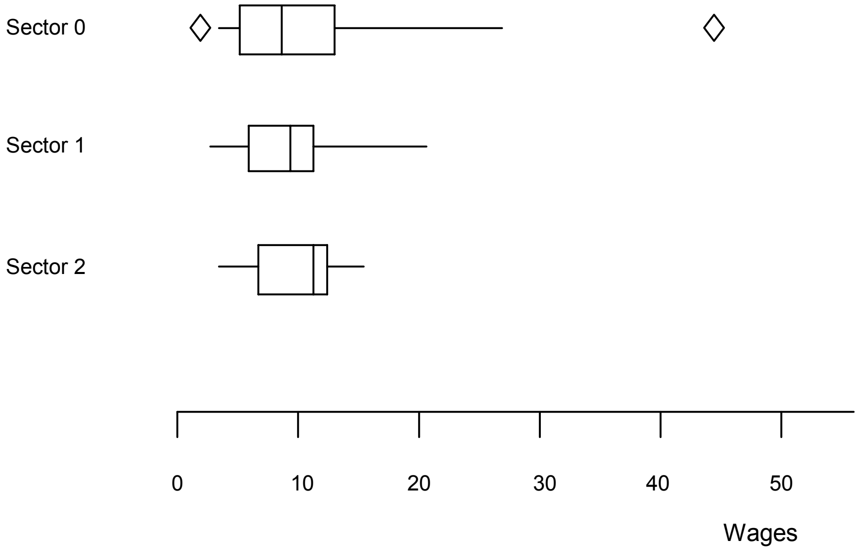
(i) A trimmed mean is likely to have removed any major outliers, and in the case of a skew distribution it will be a better central measure than the mean of all the data. Although hypothesis testing is still very approximate, descriptive statistics are improved.

(ii) These particular sets of wage data appear to be skew, as would be anticipated, rather than having many obvious outliers. In sector 0, the 44.50 and perhaps the 1.75 appear to be outliers, but there is doubt about regarding any others as such. The other noticeable "gap" is between 16.42 and 19.00 in sector 1; some people would argue for regarding the top three values in that sector as outliers.

One convention is to regard as outliers any observations below $Q_1 - 1.5R$ or above $Q_3 + 1.5R$, where R is the interquartile range ($R = Q_3 - Q_1$). This would cover all above 22.4 in sector 0, i.e. the top nine; only 20.4 in sector 1 (in spite of the obvious "gap" already mentioned); and none in sector 2.

The lowest value in sector 0 is suspect, but in a distribution of this shape no automatic calculation is likely to declare it to be an outlier.

In the boxplots, the "whiskers" have been extended all the way to the minimum and maximum for each sector except for the lowest and highest values in sector 0.



[Note. The limits of electronic reproduction may mean that the boxplots will not appear in their correct locations with precise accuracy.]

Continued on next page

(iii) Most of the available data are for sector 0. The extreme values for this sector are under $2 and over $40, but there is only one very small and one very large value. The median is below that for the other two sectors, indicating a general tendency towards lower payments. The overall pattern is skew.

Wages in sectors 1 and 2 show a rather similar pattern, but this is based on a much larger sample of data for sector 1 than for sector 2. In sector 1, there are top values (three of them) around $19 or $20, and three of $4 or less. The three top values could be checked to see if they are indeed from this sector, or perhaps from a distinctly different sub-sector compared with the rest.

Wages in sector 2 do not exceed $15 in these (few) data, but there are none below $3.75. In fact there are only three below $7. This suggests that workers are on the whole better paid at the lower end of this sector but wages do not rise to the level of the other sectors.

Higher Certificate, Paper III, 2004. Question 8

(i) Observed and expected (on the null hypothesis of no association between class of degree and sex) frequencies, and the individual contributions of each cell to the X^2 statistic, are as follows.

18	17	21.26	13.74	0.4999	0.7735
90	50	85.03	54.97	0.2905	0.4494
12	18	18.22	11.78	2.1234	3.2842
61	32	56.49	36.51	0.3601	0.5571

The test statistic is $X^2 = \sum \frac{(O-E)^2}{E}$ = 8.34. Refer this to χ_3^2. The upper 5% point is 7.815, so the result is significant at the 5% level. We have evidence to reject the null hypothesis – it seems there is a relation.

The individual contributions to X^2 show that the main contributions come from the cells for the 3rd class degree (the third row of the table), where we find fewer males and more females than would be expected. This is also the case for 1st class degrees, balanced by the opposite being true for 2nd class and Pass degrees, but these cells do not make such a marked contribution. These comments would be the substance of the report.

(iii) We have $\hat{p}$ = 35/298 = 0.117 for this organisation. The national population value of p is 0.083. We want to test the null hypothesis that the proportion in this organisation is the same as the national value, against the alternative that this organisation's value is higher. With a sample of size as large as 298, even with p as small as 0.083, a Normal approximation should be adequate. So the test statistic (without continuity correction) is

$$\frac{\hat{p}-p}{\sqrt{p(1-p)/n}} = \frac{0.117-0.083}{\sqrt{(0.083)(0.917)/298}} = \frac{0.034}{0.016} = 2.13,$$

which is referred to N(0, 1) in a one-sided interpretation. The result is significant at the 5% level (critical point 1.645) and approaching significance at the 1% level (critical point 2.326). There is considerable evidence that this organisation's proportion is higher than the national value.

The proportions of males and females with first class degrees are likely to be different, so to pool all the data into a single binomial distribution and test is not strictly correct.

Continued on next page

(iii) For this organisation, $\hat{p}_M = 90/181 = 0.497$, $\hat{p}_F = 50/117 = 0.427$. The estimated variance of $\hat{p}_M - \hat{p}_F$ is given by

$$\frac{0.497 \times 0.503}{181} + \frac{0.427 \times 0.573}{117} = 0.003472 .$$

So the (Normal approximation) test statistic for testing the null hypothesis that the true values of p_M and p_F are equal is

$$\frac{0.497 - 0.427(-0)}{\sqrt{0.003472}} = 1.19 .$$

Referring this to N(0, 1), the result is not significant – there is no evidence to suggest that p_M and p_F are not equal.

An alternative method for this part is a 2 × 2 contingency table. The observed frequencies are

	2nd class	*Other*
Male	90	91
Female	50	67

and the expected frequencies if there is no association are

	2nd class	*Other*
Male	85.034	95.966
Female	54.966	62.034

These give X^2 test statistic 1.39 (without use of Yates' correction) which, on reference to χ_1^2, is not significant.