THE ROYAL STATISTICAL SOCIETY

2002 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, 2002. Question 1

(i) If two independent sets of data are available from Normally distributed populations of measurements, with possibly different means μ_1, μ_2 but the same variance σ^2, a two-sample t test is used. It compares the sample means, $\bar{x}_1$ and $\bar{x}_2$, based on n_1 and n_2 observations, against the null hypothesis $\mu_1 = \mu_2$.

The test statistic is

$$\frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}},$$ to be referred to the $t_{n_1+n_2-2}$ distribution, and where

$$s^2 = \frac{\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)^2 + \sum_{j=1}^{n_2}(x_{2j} - \bar{x}_2)^2}{n_1 + n_2 - 2}$$ is a "pooled" estimate of σ^2.

This is not valid when the data sets have been collected from the same units (and so are not independent). In this case there will be pairs of data (x_{1i}, x_{2i}); for example in a medical trial these could be blood pressures before and after a standard exercise programme, or levels of a chemical before and after treatment with a drug. Each patient is now acting as his or her own "control" and systematic patient-to-patient variation is removed. The measurement for analysis is $d_i = x_{2i} - x_{1i}$ for each pair ($i = 1, 2, \ldots, n$). A one-sample test of "$\mu_d = 0$" is now appropriate, using t_{n-1}. This procedure is called a paired test.

(ii) A two-sample test is required, and since the chickens were all of similar age a common value of σ^2 may be assumed.

$n_1 = n_2 = 12$.

$\sum x_1 = 160$, $\sum x_1^2 = 2196$, $\bar{x}_1 = 13.33$, $s_1^2 = 5.6970$.

$\sum x_2 = 184$, $\sum x_2^2 = 2870$, $\bar{x}_2 = 15.33$, $s_2^2 = 4.4242$.

$s^2 = \frac{1}{22}\left(11s_1^2 + 11s_2^2\right) = 5.0606$.

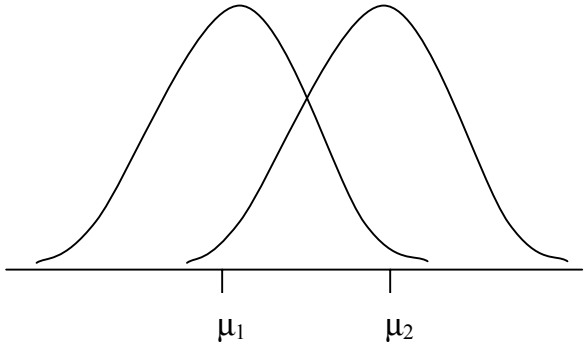
The test statistic is $\frac{13.33 - 15.33}{\sqrt{5.0606\left(\frac{1}{12} + \frac{1}{12}\right)}} = -\frac{2.00}{0.918} = -2.178$.

This is significant at the 5% level (the two-tailed 5% point of t_{22} is 2.074).

The null hypothesis $\mu_1 = \mu_2$ is rejected at the 5% level, and so there is evidence that the two food regimes differ in effect.

Higher Certificate, Paper II, 2002. Question 2

(a) In a simple significance test, there is a null hypothesis (NH or H_0) which is the basis for calculations and an alternative hypothesis (AH or H_1) which is accepted when the NH is rejected. For example, NH may be that data come from $N(\mu_1, \sigma^2)$ and AH that they come from $N(\mu_2, \sigma^2)$, with $\mu_2 > \mu_1$.



(i) Type I error = P(reject H_0 when H_0 is true).

(ii) Type II error = P(<u>not</u> reject H_0 when H_1 is true).

(iii) Level of significance = P(Type I error) = α.

(iv) Power = $1 - \beta$, where $\beta = P$(Type II error).

(b) (i) $n = 9$. $\bar{x} = 222.0$. $s^2 = 23.50$. $H_0 : \mu = 220$; $H_1 : \mu > 220$.

Test statistic is $\dfrac{222.0 - 220.0}{\sqrt{23.50/9}} = \dfrac{2.0}{1.616} = 1.238$, refer to t_8.

A one-tail test is required, so the 5% point of t_8 is 1.860. The result is not significant. There is no evidence that the recommended intake is exceeded, on average.

(ii) If n = 25, with the same values of $\bar{x}$ and s^2 as in (i), the test statistic is $\dfrac{2.0}{\sqrt{23.50/25}} = 2.063$ which is referred to t_{24}. This is significant as a one-tail test (5% point 1.711). Therefore we may reject the null hypothesis and accept that the recommended intake <u>is</u> exceeded. A larger sample size has given a more powerful test.

Higher Certificate, Paper II, 2002. Question 3

Major points to be included are as follows.

(1) Percentages of GDP for Exports and Imports show a similar pattern, first increasing then reducing again. The following figures can be plotted as a time-series graph:

		1992	1993	1994	1995	1996	1997	1998	1999
%	*E*/GDP	23.56	25.35	26.38	28.35	29.14	28.47	26.47	25.77
	I/GDP	24.77	26.40	27.06	28.74	29.69	28.41	27.41	27.48
Deficit	$I - E$	1.21	1.05	0.68	0.39	0.55	–0.06	0.94	1.71

If plotted together, the pattern in ($I - E$) can also be seen, a decrease followed by an increase which was quite sharp in 1998 and 1999.

(2) Pie charts for some years, perhaps just 1992 and 1999, could be used to show the percentages of Exports and Imports which went to different regions. Percentages for 1992, 1995, 1996, 1999 (in case a year in the middle is used also) are

EXPORTS	1992	1995	1996	1999
EU	54.2	52.9	52.1	52.8
NA	16.4	15.7	16.0	19.4
Other	29.4	31.3	31.9	27.8

IMPORTS	1992	1995	1996	1999
EU	55.5	54.7	53.5	53.1
NA	14.4	15.2	15.7	16.3
Other	30.1	30.1	30.7	30.6

(3) Indices of 1999 relative to 1992 could be calculated (1992 = 100):

Exports:	EU	156.1	Imports:	EU	155.7
	NA	188.8		NA	183.4
	Other	151.9		Other	165.1
	Total	160.3		Total	162.5

Note that current prices are used, whereas scaling to constant prices is more helpful in understanding changes.

Higher Certificate, Paper II, 2002. Question 4

(a) ***[Note.*** *The discussion presented here includes aspects that are further explored in Higher Certificate Paper III and/or in the Graduate Diploma Applied Statistics papers.]*

In two-way analysis of variance, there are two "factors" (A and B) that are sources of systematic variation that might affect the outcome. Factor A has a "levels" and factor B has b "levels". The observation when A is at level i and B is at level j is denoted by y_{ij} (for the time being we suppose there is only one observation at each such combination). The usual linear model is that an observation y_{ij} can be explained in terms of an overall mean μ, an effect (α_i, deviation from μ) due to having the ith level of A, similarly an effect (β_j, deviation from μ) due to having the jth level of B, and a "residual" term ε_{ij} which explains the random natural variation and is assumed N(0, σ^2) where σ^2 is constant for all observations. Thus the model is

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \qquad i = 1, 2, ..., a \quad j = 1, 2, ..., b\,.$$

This is often applied for designed experiments where one of the factors is a "blocking factor" and the other represents the "treatments" that are actually being compared. The factors are often then called "blocks" and "treatments", with appropriate Greek letters (β and τ) being used.

Sometimes there is more than one observation ("replication") for each combination of a level from factor A and a level from factor B – usually the same number of observations, say n, for every combination – and in such cases an observation is denoted by y_{ijk} where the third subscript k (k = 1, 2, …, n) indicates the replicate. The residual term correspondingly needs to be denoted by ε_{ijk}. The model then becomes

$$y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk} \qquad i = 1, 2, ..., a \quad j = 1, 2, ..., b \quad k = 1, 2, ..., n$$

and can be further extended to include also an "interaction" term, usually denoted by $(\alpha\beta)_{ij}$. This represents the situation where some levels of A give a 'better' result in combination with some levels of B whereas other levels of A give a 'better' result with other levels of B. It is often assumed that there is no such interaction, especially in the blocks-and-treatments designed experiments situation.

(b) Totals (of rows and columns in the given table) are as follows.

Gravel types (i = 1, 2, 3):

	1	*2*	*3*	Grand total
Total	46	57	86	189
Mean	15.33	19.00	28.67	

Cement types (j = 1, 2, 3, 4 for A, B, C, D respectively):

	A	*B*	*C*	*D*	Grand total
Total	42	49	60	38	189
Mean	14.00	16.33	20.00	12.67	

Continued on next page

Sum of squares for gravel = $\frac{1}{4}\left(46^2+57^2+86^2\right)-\frac{189^2}{12}=3190.25-2976.75$

= 213.50.

Sum of squares for cement = $\frac{1}{3}\left(42^2+49^2+60^2+38^2\right)-\frac{189^2}{12}=3069.667-2976.750$

= 92.917.

Total sum of squares = $3293-\frac{189^2}{12}=316.25$. [3293 is $\Sigma\Sigma y_{ij}^2$.]

Residual sum of squares (obtained by subtraction) = 316.25 – 92.917 – 213.50

= 9.833.

Analysis of Variance:

Source of variation	df	Sum of Squares	Mean Square	*F* value
Gravels	2	213.500	106.75	65.13 (very highly sig)
Cements	3	92.917	30.97	18.90 (highly sig)
Error (Residual)	6	9.833	1.639	
Total	11	316.250		

The *F* value of 65.13 is referred to $F_{2,6.}$ It is very much larger than even the upper 0.1% point (which is 27.00). The *F* value of 18.90 is referred to $F_{3,6}$. It substantially exceeds the upper 1% point (9.78) but not the upper 0.1% point (23.70).

The residual mean square (1.639) is the estimate of experimental error. This measures the underlying variability of the production process; it is not very large, suggesting that the process appears to be in control.

There are significant differences among both cements and gravels. The company should use the cement which gives the greatest strength, and on the evidence of these results that is clearly *C*. These are also noticeable differences among gravels; in particular, type 3 gave stronger beams than the others. In fact the differences among gravels were greater than those among cements; they would be worth exploring further, and in production this is an important factor to control.

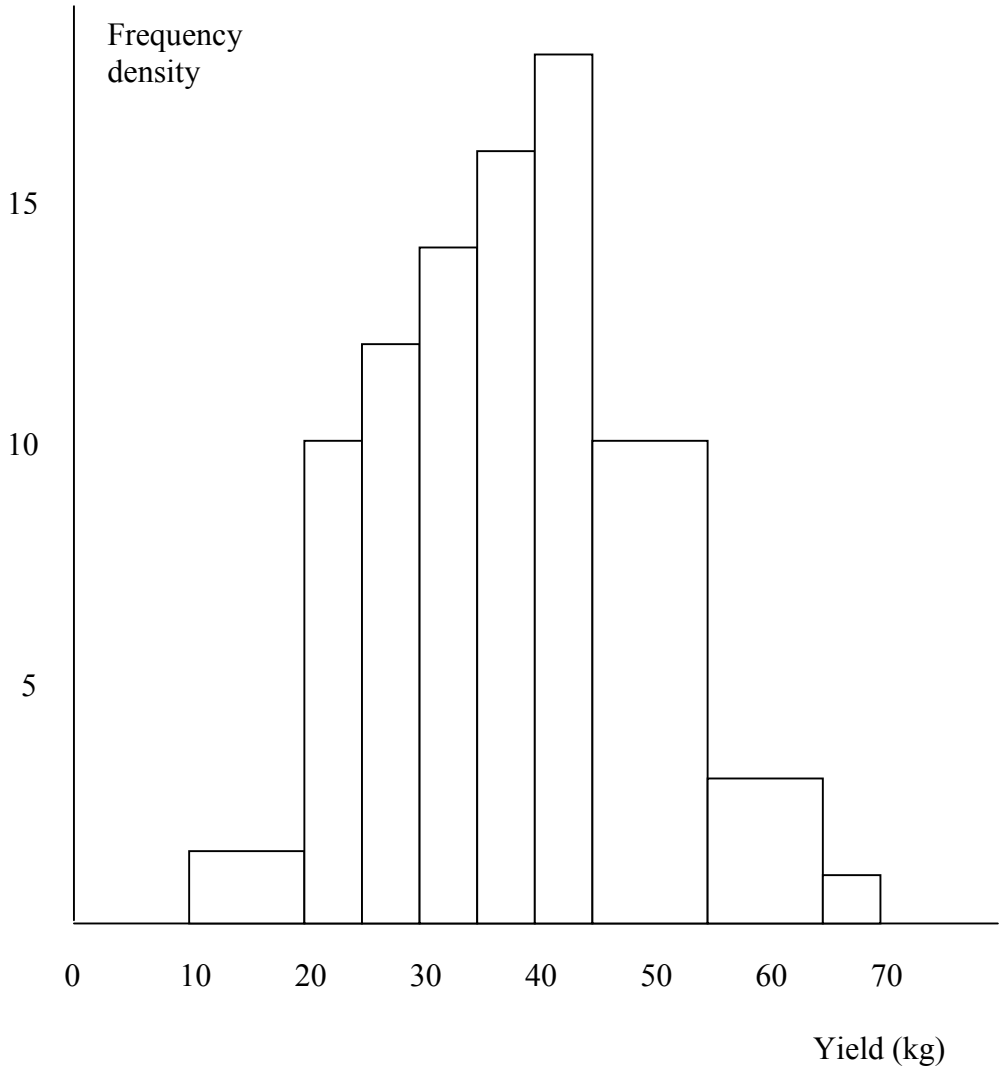
Technical back-up to go in an appendix to a report would include the significant differences between pairs of means for the cements, which are as follows (1.639 is the residual mean square, from the analysis of variance above):

$$t_6\sqrt{\frac{2\times1.639}{3}}=\begin{Bmatrix}2.447\\3.707\\5.959\end{Bmatrix}\times1.045=\begin{Bmatrix}2.56 & \text{at 5\%}\\3.87 & \text{at 1\%}\\6.23 & \text{at 0.1\%}\end{Bmatrix},$$

showing that *C* > all others at 5% or more, otherwise the only evidence of real difference is between *B* and *D* [in fact it might be reasonable to assume that *A*, *B* and *D* are all the same but *C* distinctly better]. A similar analysis for significant differences between pairs of means for the gravels would also be included.

(i)

The frequency densities in the first 8 intervals are 1.5, 10, 12, 14, 16, 18, 10 and 3 respectively. The width of the last interval must be chosen arbitrarily; it could be to 75 (or even more), but as the frequency above 55 drops off sharply it seems reasonable to assume that the last interval ends at 70. This affects the calculations in part (ii) very little, and the histogram not much. On this basis, the frequency density in the last interval is 1.



Continued on next page

(ii) Modal class is [≥ 40 but < 45], as it has greatest frequency density (allowing for different widths of intervals).

The median is the "$50\frac{1}{2}$'th" observation in ascending order.

This is at $35+\dfrac{11.5}{16}\times 5 = 38.6$.

Yield y	Frequency f	Mid-point x	fx	fx^2	Cum freq F
$10 \le y < 20$	3	15	45	675	3
$20 \le y < 25$	10	22.5	225	5062.50	13
$25 \le y < 30$	12	27.5	330	9075	25
$30 \le y < 35$	14	32.5	455	14787.50	39
$35 \le y < 40$	16	37.5	600	22500	55
$40 \le y < 45$	18	42.5	765	32512.50	73
$45 \le y < 55$	20	50	1000	50000	93
$55 \le y < 65$	6	60	360	21600	99
$y \ge 65$	1	(67.5)	67.5	4556.25	100
			3847.5	**160768.75**	

$\bar{x} = \dfrac{\sum fx}{\sum f} = 38.475$, or 38.5 to a reasonable level of accuracy.

$s^2 = \dfrac{1}{99}\left(160768.75 - \dfrac{3847.5^2}{100}\right) = \dfrac{12736.1875}{99} = 128.648$,

so the standard deviation is 11.34.

(iii) In order to calculate $\bar{x}$, all the frequency in each interval had to be concentrated at the centre. This has given an over-estimate, so there must have been more left-of-centre observations in some (or all) intervals. The median is also an over-estimate, due to assuming a uniform spread of the data in the interval 35 – 40. This also suggests some skewness in intervals.

Higher Certificate, Paper II, 2002. Question 6

(i) We might reasonably suppose that each trial (i.e. a rat trying a door) has the same probability p of success, independently of all other trials. These trials continue until there is success, on the xth trial; there is only one order in which this can occur, namely $x - 1$ failures followed by one success, so we have

$P(X = x) = (1-p)^{x-1}\, p$.

The possible values of x are 1, 2, 3, … .

In this case, if the rat is "guessing" there will be probability 1/3 of choosing the food door on any trial; i.e. p is 1/3. The weakness in this argument may be that the rat does not "guess" because it can detect food, e.g. by smell. If that occurs, P(food door) is >1/3, and unknown.

(ii) Expected frequencies are $50\left(\frac{1}{3}\right)\left(\frac{2}{3}\right)^{x-1}$ for x = 1, 2, … on the null hypothesis of a geometric distribution with p =1/3.

x	1	2	3	4	5	6	7	≥ 8	TOTAL
Obs	15	11	7	6	5	4	2	0	50
Exp	16.67	11.11	7.41	4.94	3.29	2.19	1.46	2.93	
					5.48		4.39		

Because the geometric distribution tails off very slowly, it is not easy to combine expected values, but the above grouping is better than combining the whole tail from (say) 5 upwards because the pattern is better preserved (and degrees of freedom are saved).

No parameters were estimated, so there will be 5 degrees of freedom for the usual chi-squared test using x = 1; 2; 3; 4; (5, 6); ≥ 7.

The test statistic is

$$\frac{(15-16.67)^2}{16.67}+\frac{(11-11.11)^2}{11.11}+\frac{(7-7.41)^2}{7.41}+\frac{(6-4.94)^2}{4.94}+\frac{(9-5.48)^2}{5.48}+\frac{(2-4.39)^2}{4.39}$$

= 3.98.

This value is not significant when compared with χ^2_5. The geometric hypothesis is not rejected. Therefore we may assume the model is satisfactory, and the animals do appear to be "guessing".

Higher Certificate, Paper II, 2002. Question 7

(i) When there is doubt about what distribution may be used to explain a set of data, especially when it is not possible to assume Normality (even after a transformation), methods that do not depend on distributional assumptions are useful. There are several methods for analysing sets of data using distribution-free ("non-parametric") tests, although they are less powerful than those using distribution theory when the underlying distributions are (at least approximately) Normal, so sample sizes need to be larger for non-parametric tests.

(ii) (a) If there has been no effect, the number of patients who lose weight should be binomially distributed, $n = 14$, $p = ½$. A sign test allocates (say) + sign to those who have lost weight and – sign to those who have not, and does not use any whose weights remain exactly the same. There are 11 + signs, out of 14.

If B(14, ½) explains the situation, we have

$$P(11,12,13 \text{ or } 14 \text{ plus (+) signs}) = \left(\frac{1}{2}\right)^{14}\left(\binom{14}{11}+\binom{14}{12}+\binom{14}{13}+\binom{14}{14}\right)$$

$$= \left(\frac{1}{2}\right)^{14}\left(\frac{14.13.12}{3.2.1}+\frac{14.13}{2.1}+14+1\right)$$

$$= \left(\frac{1}{2}\right)^{14}(364+91+15) = \frac{470}{2^{14}} = 0.0287.$$

A two-tail test using null hypothesis "no effect" and alternative hypothesis "some effect" (unspecified) therefore has *p*-value 0.0574, and does not provide evidence on which to reject the null hypothesis.

(b) A Wilcoxon signed-rank test uses the sizes as well as the signs of the differences, and so carries more power to reject the null hypothesis when it is false.

Patient	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Difference	60	44	10	-1	-3	5	52	-8	14	10	15	31	14	3
Rank	14	12	6.5	1	2.5	4	13	5	8.5	6.5	10	11	8.5	2.5

The ranks are those of the absolute differences. The sums of the ranks for the positive and negative differences are $T_+ = 96\frac{1}{2}$ and $T_- = 8\frac{1}{2}$. The test statistic is $T = \min(T_-, T_+) = 8.5$. The tables, with $n = 14$ and $\alpha = 0.05$ (two-sided), give $T = 21$. Since 8.5 is (much) lower than this, there is evidence to reject the null hypothesis. Inspection of the data shows that the negatives (i.e. weight gains) are generally small in size compared with the positives (weight losses).

Higher Certificate, Paper II, 2002. Question 8

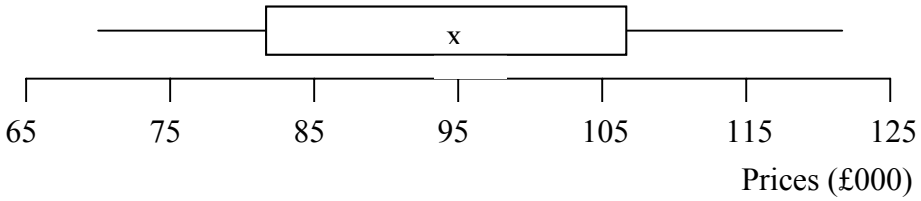
(i) Prices (£000) in rank order are

68, 74.95, 75, 78, 79.95, 82.95, 85, 85, 85.95, 95, 95, 95, 97.5, 99.95, 105, 108, 115, 119.95, 120, 122.95.

The median is the "$10\frac{1}{2}$'th" observation in ascending order. The 10th and 11th are both 95, so the median is 95.

The lower quartile is between the 5th and 6th, which are 79.95 and 82.95. It is acceptable to take the average of these, though other detailed definitions are also in use. The upper quartile is found in a similar way. On this basis, we take

lower quartile: $\frac{1}{2}(79.95+82.95)=81.45$, upper quartile: $\frac{1}{2}(105+108)=106.5$.



*[**Note.** This plot might not appear **exactly** correct, due to screen and/or printer resolution.]*

The median is approximately in the middle of the box, and the two whiskers are about the same length, so the distribution is not far from symmetrical. There is also some clustering in the middle, near to the median, so a Normal distribution could be proposed as a model for these data.

Continued on next page

(ii) For these data (using a pocket calculator; some care is needed to preserve accuracy in dealing with the large numbers both here and for the adjacent suburb), $\sum y = 1\,888\,150$ and $\sum y^2 = 183\,450\,567\,500$, giving s = 16535.5222, so $s_y^2 = 273\,423\,493.4211$. ($n = 20$; 19 degrees of freedom.)

For the adjacent suburb, n = 30, $\sum x = 2864490$ and $\sum x^2 = 278\,338\,961\,408$, so $s_x^2 = \frac{1}{n-1}\left\{\sum x^2 - \frac{\left(\sum x\right)^2}{n}\right\}$ is $\frac{1}{29}\left(4\,828\,862\,738\right) = 166\,512\,508.2069$. (29 degrees of freedom.)

Test statistic for equality of variances is $s_y^{\,2} / s_x^{\,2}$ = 1.642 which is not significant when referred to $F_{19,29}$. There is not sufficient evidence to reject the null hypothesis, which is "$\sigma_X^2 = \sigma_Y^2$".

The F test is valid on the basis of apparent Normality for the original data, and the assumption that this is also true for the other sample.