THE ROYAL STATISTICAL SOCIETY

2006 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.

Note. In accordance with the convention used in the Society's examination papers, the notation log denotes logarithm to base *e*. Logarithms to any other base are explicitly identified, e.g. $\log_{10}$.

Higher Certificate, Paper II, 2006. Question 1

(i) The variance of this sample is

$$s^2 = \frac{1}{59}\left(4054484 - \frac{15568^2}{60}\right) = \frac{15106.93}{59} = 256.05 .$$

The null hypothesis to be tested is "$\sigma^2 = 256$". It seems obvious that this null hypothesis is not likely to be rejected (even if the sample had been of considerably smaller size), but continuing with a formal test we use test statistic

$$\frac{(n-1)s^2}{\sigma^2} = \frac{59 \times 256.05}{256} = 59.01$$

which is referred to χ^2_{59}. The upper 5% point is about 78. Clearly we cannot reject the null hypothesis as the data give no evidence for doing so.

(ii) We have $\bar{x} = \frac{15568}{60} = 259.46$ and we wish the test the null hypothesis μ = 266. Taking the value of σ as 16, which seems highly plausible from part (i), we use test statistic

$$\frac{\bar{x} - 266}{\frac{16}{\sqrt{60}}} = -3.16$$

and refer to N(0, 1).

[**Alternatively,** we could continue to use the sample variance s^2 (= 256.05) and refer $\frac{\bar{x} - 266}{\frac{s}{\sqrt{60}}}$ to t_{59}; this makes hardly any difference in practice in this case.]

This is well beyond the double-tailed 1% point of N(0, 1); there is strong evidence against this null hypothesis. It is reasonable to conclude that this population has a mean different from the "normal" one; it appears to be less.

Using N(0, 1), we have $\Phi(-3.16) = 0.0008$, giving a p-value of 0.0016.

Higher Certificate, Paper II, 2006. Question 2

(i) Parametric tests need assumptions about the distribution underlying the data – often that it is Normal (if the situation is continuous). Data in the form of subjective scores are unlikely to follow any of the common distributions, and two sets of independent data may not even be of the same shape, location, scatter or skewness. Non-parametric tests allow simple characteristics of distributions to be compared with few or no theoretical assumptions. However, they have less power than corresponding parametric tests in cases where the parametric test is in fact valid. They therefore need larger sample sizes.

(ii) The Wilcoxon rank sum test (or, equivalently, the Mann Whitney U form of this test) is suitable for this comparison. First rank all 20 data items, as follows.

Data	0	3	8	9	10	14	15	16	17	22	28	31	33	37
Ranks	1	2	3	4	5	6	7	8	9	10	11	12	13	14
E or *A*	*E*	*E*	*A*	*E*	*E*	*E*	*E*	*A*	*E*	*E*	*A*	*A*	*A*	*A*

38	39	47	50	55	80
15	16	17	18	19	20
A	*A*	*E*	*A*	*A*	*E*

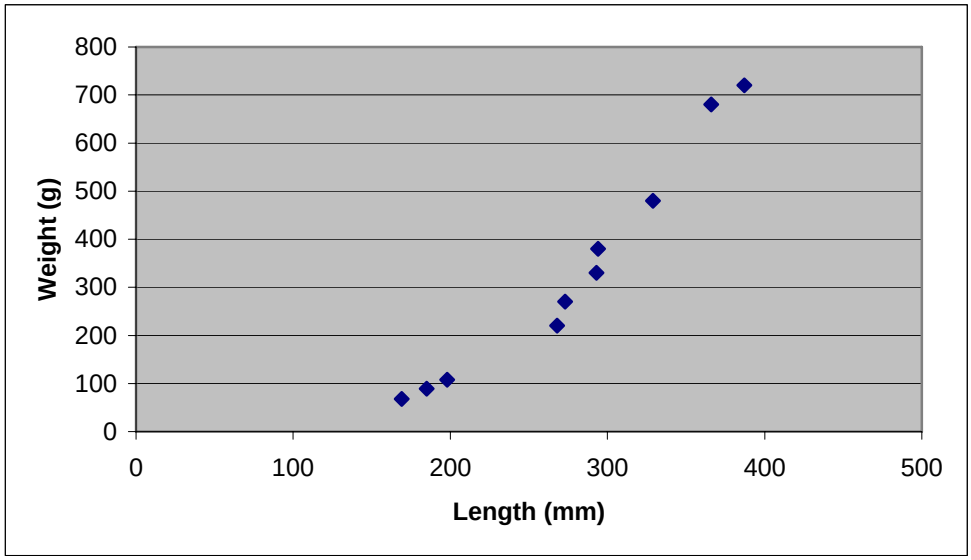
The rank sum for Entonox (E) is 1 + 2 + 4 + 5 + 6 + 7 + 9 + 10 + 17 + 20 = 81. That for A is 129.

The required test is two-sided. For a 5% test, we refer the smaller of these (81) to the lower 2½% point for the $W_{10,10}$ distribution as shown in the Society's statistical tables for use in examinations. This is 78 so, at the 5% level of significance, we cannot reject the null hypothesis that pain scores do not differ. However, we note that the result is (just) significant at the 10% level (the lower 5% point is 82), and the sample sizes are quite small. So, overall, we do not really have sufficient evidence to say whether or not there is an advantage for Entonox. A more powerful test should be conducted using larger samples before coming to a firm decision.

The data do appear to need a non-parametric testing procedure.

Higher Certificate, Paper II, 2006. Question 3

(i) The graph suggests that a linear fit will be reasonable, at least as a first approximation. There may be curvature in the relation of weight and length.



(ii) (a) $\hat{\beta}_1 = \dfrac{S_{xy}}{S_{xx}}$.

$$S_{xy} = \sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n} = 1075861 - \frac{2762 \times 3345}{10} = 151972.0.$$

$$S_{xx} = 812594 - 2762^2/10 = 49729.6.$$

$$\therefore \hat{\beta}_1 = 3.056.$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 334.5 - (3.056 \times 276.2) = -509.56.$$

(b) The first three points are not fitted at all well. Note that the intercept is –509.56, whereas it looks as if it should be much nearer 0 if these are to be fitted. But the remaining points are fitted reasonably well. It would however be sensible to examine also a quadratic relationship.

(iii) The coefficient of determination is given by $R^2 = S_{xy}^2 / S_{xx} S_{yy}$.

$$S_{yy} = 1610009 - \frac{3345^2}{10} = 491106.5.$$

$$\therefore R^2 = \frac{151972.0^2}{49729.6 \times 491106.5} = 0.9457.$$

Thus 94.6% of the total variation in the weights of the sea bass is explained by a linear relationship with their lengths.

Higher Certificate, Paper II, 2006. Question 4

(i) The key assumption is that the experimental errors are independent N(0, σ^2) variables (note constant σ^2).

Totals are

Method 1	Method 2	Method 3
78.6	79.3	86.9

Subj. 1	Subj. 2	Subj. 3	Subj. 4	Subj. 5	Subj. 6	Subj. 7	Subj. 8
23.2	30.8	26.8	26.5	32.9	39.5	27.2	37.9

The grand total is 244.8. $\Sigma\Sigma y_{ij}^2 = 2585.22$.

"Correction factor" is $\frac{244.8^2}{24} = 2496.96$.

Therefore total SS = 2585.22 – 2496.96 = 88.26.

SS for methods = $\frac{78.6^2}{8} + \frac{79.3^2}{8} + \frac{86.9^2}{8} - 2496.96 = 5.30$.

SS for subjects = $\frac{23.2^2}{3} + \frac{30.8^2}{3} + \ldots + \frac{37.9^2}{3} - 2496.96 = 78.47$.

The residual SS is obtained by subtraction.

Hence the analysis of variance table is as follows.

SOURCE	DF	SS	MS	F value
Methods	2	5.30	2.65	8.26 Compare $F_{2,14}$
Subjects	7	78.47	11.21	34.95 Compare $F_{7,14}$
Residual	14	4.49	0.3207	$= \hat{\sigma}^2$
TOTAL	23	88.26		

Upper critical points of $F_{2,14}$ and $F_{7,14}$ are as follows.

	5%	1%	0.1%
$F_{2,14}$	3.74	6.51	11.78
$F_{7,14}$	2.76	4.28	7.08

Solution continued on next page

The F value for methods is highly significant; we have strong evidence that not all the methods are the same in terms of mean clotting time. The F value for subjects is very highly significant. We have very strong evidence that not all the subjects are the same in this regard; the analysis has detected and removed a large systematic source of variation.

To investigate method differences, we need the method means, which are

Method 1 : 9.825 Method 2 : 9.9125 Method 3 : 10.8625.

The least significant difference between any pair of these means is

$$t_{14}\sqrt{\frac{2\times 0.3207}{8}} = 0.283\, t_{14} \quad \text{where} \quad t_{14} = \begin{cases} 2.145 & \text{at } 5\% \\ 2.977 & \text{at } 1\% \\ 4.140 & \text{at } 0.1\% \end{cases}$$

so the least significant differences are 0.607 for 5%, 0.842 for 1% and 1.172 for 0.1%. Clearly methods 1 and 2 do not appear to differ in mean clotting time but there is strong evidence that method 3 has a higher mean clotting time than either of the others.

(ii) In the analysis of variance now, there will be no "subjects" term, only "methods" and "residual". The new residual will include *both* the amount previously classified as residual *and* the amount previously classified as the subjects term. Thus the new analysis of variance table is as follows.

SOURCE	DF	SS	MS	F value
Methods	2	5.30	2.65	0.67 Compare $F_{2,21}$
Residual	21	82.96	3.95	$= \hat{\sigma}^2$
TOTAL	23	88.26		

The F value for methods is now not significant – we have no evidence to reject the null hypothesis that the methods are the same in terms of mean clotting time. This is because the apparent underlying variability in the data is now very high, due to the consistent but unidentified between-subject variation. Blocking therefore greatly increased the power to detect differences between the methods.

Higher Certificate, Paper II, 2006. Question 5

(i) Ordered diagram: (stem unit 10000)

STEM	
0	1 1 2 3 3 4 4 6 7 8 8 9 9
1	4 4 7
2	1 4 4 7 8
3	3 6 6 7 8 9
4	2 6 9
5	0 0 2 7 7
6	1
7	0 1 3 8 9
8	2 7 8
9	3 6 9
10	9
11	5
…	
16	0

There is considerable skewness, with a large number in stem 0 and a long tail to the right. There are also gaps.

(ii) The median is between the 25th and 26th in order: $M = \frac{37+38}{2} = 37.5$.

The quartiles are at the 13th and 38th in order: lower quartile q = 9, upper quartile Q = 71. [Other conventions are also acceptable for the quartiles.]

These are in thousands; so we have q = 9000, M = 37500, Q = 71000. The inter-quartile range is then 62000.

$\overline{x} = \frac{2217}{50} = 44.34$ (thousand).

$s^2 = \frac{1}{49}\left(164441 - \frac{2217^2}{50}\right) = \frac{66139.22}{49} = 1349.78$; so $s = 36.74$ (thousand).

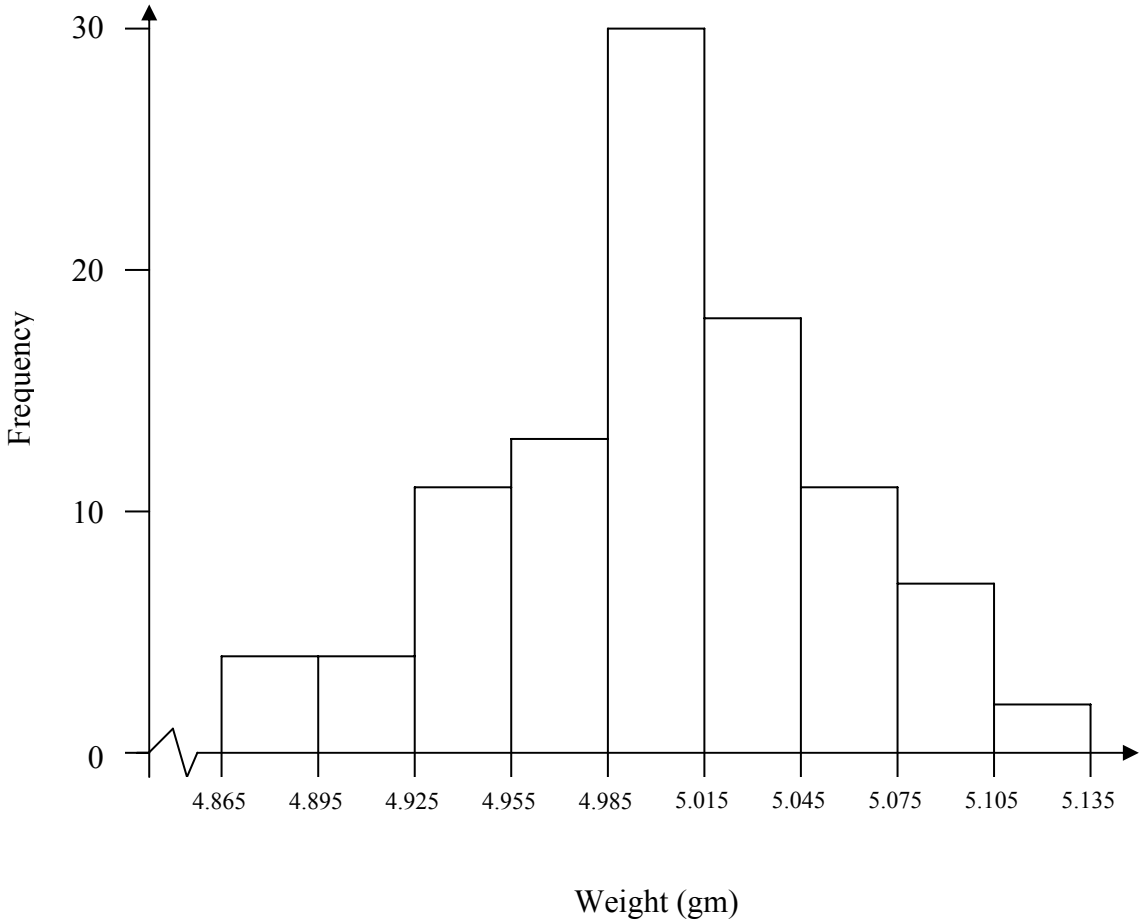
(iii) For reasons given above, the mean and standard deviation will not be good measures of location and dispersion. The median, 37500, and inter-quartile range, 62000 (or the semi-iqr 31000) would be preferred.

The middle 50% of the observations have range 62000. The lower 50% are 37500 or less.

(iv) A correlation measure is appropriate. The strength of a linear relationship can be assessed by Pearson's product-moment coefficient. But the farm number distribution is skew, and the state area distribution is also likely to be skew, so Spearman's rank-based coefficient is better. (Spearman's coefficient uses the ranked data in each set instead of actual measurements.)

Higher Certificate, Paper II, 2006. Question 6

(i) In a histogram for open-ended data, it is common to assume that the intervals at the beginning and end have the same width as others (unless there is good reason to do otherwise), and this has been done here.



(ii) Using the sample variance $s^2 = 0.055^2$, we work with $W \sim \text{N}(5, 0.055^2)$. This gives $Z = \dfrac{W-5}{0.055} \sim \text{N}(0, 1)$. The table on the next page shows the value of w at the end-point of each interval, the corresponding value of z, the probability $P(Z < z)$, and hence the probability of being in the corresponding interval. The observed and corresponding expected frequencies (o and e) are shown in the last two columns of the table.

Solution continued on next page

Weight w	$w-5$	$z=\frac{w-5}{0.055}$	$P(Z<z)$	Prob. in interval		o	e
4.895	–0.105	–1.9091	0.0281	0.0281	(from $-\infty$)	4	2.81
4.925	–0.075	–1.3636	0.0863	0.0582		4	5.82
4.955	–0.045	–0.8182	0.2066	0.1203		11	12.03
4.985	–0.015	–0.2727	0.3925	0.1859		13	18.59
5.015	0.015	0.2727	0.6075	0.2150		30	21.50
5.045	0.045	0.8182	0.7934	0.1859		18	18.59
5.075	0.075	1.3636	0.9137	0.1203		11	12.03
5.105	0.105	1.9091	0.9719	0.0582		7	5.82
				0.0281	(to $+\infty$)	2	2.81

For the test, the expected frequencies need to be not too small (≥5 is often used as a criterion). On this basis, we combine the first two cells and the last two cells to get the following table.

o	8	11	13	30	18	11	9
e	8.63	12.03	18.59	21.50	18.59	12.03	8.63

The test statistic is

$$X^2=\sum\frac{(o-e)^2}{e}=\frac{(8-8.63)^2}{8.63}+\frac{(11-12.03)^2}{12.03}+\ldots+\frac{(9-8.63)^2}{8.63}=5.30,$$

which is referred to χ^2_5 (note 5 degrees of freedom because the table has 7 cells and there is one [NB **only** one, i.e. σ^2] estimated parameter). This is not significant (the 5% point is 11.07). The null hypothesis that a Normal distribution with mean 5 (years) underlies the data cannot be rejected on this evidence.

Higher Certificate, Paper II, 2006. Question 7

(i) Independent samples are those where different (and unrelated) subjects, or units of experimental material, are used for the two samples. They are randomly selected from their corresponding populations, and a measurement is taken on each unit. For example, the units in one sample of seedlings taken from a forestry nursery are given treatment *A*, and those in the other sample treatment *B*; *A* and *B* might be different fertiliser or cultivation treatments given at the same stage in plant growth. The measured response is a size or health measurement taken at the same age of plants.

The populations underlying the sets of sample data are assumed to be Normally distributed with the same variance.

Paired samples use the same units for both treatments (e.g. a "before-and-after" study), or pairs of units as closely alike as possible. For example, a medical trial of alternative drugs for a chronic (long-lasting) condition might use pairs of patients whose conditions before treatment are very similar, one of each pair receiving drug *A* and the other drug *B* (allocated at random).

The underlying distribution of differences within pairs needs to be assumed Normal.

(ii) Let *H*, *W* represent the ages of Husband and Wife. We have 15 (paired) observations h_i, w_i and the differences d_i are

7, 7, 5, –4, –2, 4, 1, –6, 1, –2, 1, 2, 2, –3, 4.

(a) We have $\Sigma d_i = 17$, $\Sigma d_i^2 = 235$. Thus $\bar{d} = 17/15 = 1.133$ and

$$s_d^{\,2} = \frac{1}{14}\left(235 - \frac{17^2}{15}\right) = 15.4095\,,$$ giving $s_d = 3.925$.

(b) The test statistic for testing the null hypothesis that $\mu_H = \mu_W$ is

$$\frac{\bar{d} - 0}{s_d/\sqrt{15}} = \frac{1.133}{1.014} = 1.12\,,$$

which is referred to t_{14}. This is not significant, so the null hypothesis cannot be rejected. There is no evidence of a difference between the mean ages.

(c) The confidence interval is given by $\bar{d} \pm t\dfrac{s_d}{\sqrt{15}}$ where t is the double-tailed 5% point of t_{14} i.e. 2.145. Thus the interval is given by $1.133 \pm (2.145 \times 1.014)$, i.e. 1.133 ± 2.175, i.e. it is (–1.04, 3.31).

Higher Certificate, Paper II, 2006. Question 8

(i) We have a 2×2 contingency table. The null hypothesis is that there is no association between an individual's sex and the chance of he or she having a recently recorded cholesterol measurement. The contingency table is as follows, with the expected frequencies in brackets in each cell (e.g. 88.48 = $131 \times 206 / 305$).

		Cholesterol level recorded		
		No	*Yes*	Total
Sex	*Female*	109 (88.48)	22 (42.52)	131
	Male	97 (117.52)	77 (56.48)	174
	Total	206	99	305

All the differences between observed and expected frequencies are ±20.52, becoming ±20.02 if Yates' correction is used. Thus the usual test statistic can be calculated as (using Yates' correction)

$$(20.02)^2 \left\{ \frac{1}{88.48} + \frac{1}{42.52} + \frac{1}{117.52} + \frac{1}{56.48} \right\} = 24.46$$

(or 25.70 if Yates' correction is not used). This is referred to χ^2_1. This is very highly significant (for example, the 1% point is 6.635); we have very strong evidence to reject the null hypothesis and conclude that there is an association.

(ii) $p_f - p_m$ is estimated by $\hat{p}_f - \hat{p}_m = \frac{22}{131} - \frac{77}{174} = 0.168 - 0.443 = -0.275$. The estimated variance of $\hat{p}_f - \hat{p}_m$ is given by

$$\frac{\hat{p}_f (1 - \hat{p}_f)}{n_f} + \frac{\hat{p}_m (1 - \hat{p}_m)}{n_m} = 0.001067 + 0.001418 = 0.002485.$$

Thus the approximate 95% confidence interval for $p_f - p_m$ is given by –0.275 ± (1.96×√0.002485) i.e. it is (–0.177, –0.373).

(iii) There is clear evidence that the proportions are not the same for men and women. In part (i), this is interpreted via the very strong evidence of an association. In part (ii), the confidence interval does not contain 0, indeed it is a long way from 0, again giving very strong evidence of a real difference.