THE ROYAL STATISTICAL SOCIETY

2008 EXAMINATIONS – SOLUTIONS

HIGHER CERTIFICATE

(MODULAR FORMAT)

MODULE 3

BASIC 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, Module 3, 2008. Question 1

(i) Population mean = (1 + 4 + 9 + 16 + 25)/5 = 55/5 = 11.

Population variance = $\{(1^2 + \ldots + 25^2)/5\} - 11^2$ = 195.8 – 121 = 74.8.

(ii) The samples and the values of the sample mean are as follows.

Sample		
x_1	x_2	Mean $\bar{x}$
1	1	1
1	4	2.5
1	9	5
1	16	8.5
1	25	13
4	1	2.5
4	4	4
4	9	6.5
4	16	10
4	25	14.5
9	1	5
9	4	6.5
9	9	9
9	16	12.5
9	25	17
16	1	8.5
16	4	10
16	9	12.5
16	16	16
16	25	20.5
25	1	13
25	4	14.5
25	9	17
25	16	20.5
25	25	25

(iii) $E(\bar{X}) = E(X) = 11$. $\mathrm{Var}(\bar{X}) = \dfrac{\mathrm{Var}(X)}{2} = 37.4$.

[Note. These could also be worked out from first principles by first listing all the values of $\bar{x}$ and their frequencies of occurrence.]

(iv) 5 of the 25 values of $\bar{x}$ are greater than 16.5, so $P(\bar{X} > 16.5) = 5/25 = 0.2$.

The required approximating Normal distribution is N(11, 37.4); using this,

$$P(\mathrm{N}(11,37.4) > 16.5) = P\left(\mathrm{N}(0,1) > \frac{16.5-11}{\sqrt{37.4}}\right) = P(\mathrm{N}(0,1) > 0.8993)$$

= 1 – 0.8157 = 0.1843.

The values of X are skewed to the right, so the distribution of $\bar{X}$ is positively skewed. However, the Normal distribution is symmetrical, so we expect it to understate the true probabilities in the right-hand tail.

Higher Certificate, Module 3, 2008. Question 2

(i) $R_i \sim \mathrm{B}(n_i,\, p_i)$ and so, for large n_i, R_i is approximately $\mathrm{N}(n_i p_i,\, n_i p_i(1-p_i))$. Hence $R_i/n_i \sim$ (approx) $\mathrm{N}\left(p_i,\, p_i(1-p_i)/n_i\right)$.

$$\therefore D \sim \text{(approx) } \mathrm{N}\left(p_1 - p_2,\ \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right).$$

For the null case $p_1 = p_2 = p$, we have the approximation

$$D \sim \mathrm{N}\left(0,\ p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$$

and thus

$$Z = \frac{D}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim \mathrm{N}(0,1), \text{ approximately.}$$

Therefore an approximate size-α test of the null hypothesis of common (unspecified) proportion p, against the two-sided alternative that the proportions differ, is given by finding the observed value z of Z and rejecting the null hypothesis if $|z| > z_{\alpha/2}$, where $z_{\alpha/2}$ is the value such that a N(0, 1) random variable exceeds it with probability $\alpha/2$ (e.g. 1.96 for $\alpha = 0.05$). A value for p is needed in the denominator of z; as the samples are large, the approximate test will remain valid if p is replaced by its combined-samples estimate $\hat{p}$.

(ii) $n_1 = n_2 = 1000$; $\frac{1}{n_1} + \frac{1}{n_2} = 0.002$; $\frac{r_1}{n_1} - \frac{r_2}{n_2} = 0.08 - 0.006 = 0.074$.

$\hat{p} = (86/2000) = 0.043$.

So $z = \dfrac{0.074}{\sqrt{0.043 \times 0.957 \times 0.002}} = 8.157$.

This is very much larger than $z_{0.0005}$ (which is 3.2905). So, even at this stringent level of significance, the null hypothesis of equality of proportions is decisively rejected. There is extremely strong evidence that colour-blindness proportions differ between males and females (it appears to be more prevalent among males).

Solution continued on next page

(iii) For the confidence interval, we need the approximate standard deviation of D with separate p_1 and p_2. Using the separate sample estimates of p_1 and p_2, this is

$$\sqrt{\frac{0.08\times 0.92}{1000}+\frac{0.006\times 0.994}{1000}} = 0.0089.$$

Hence the required approximate 95% confidence interval is given by $0.074 \pm (1.96\times 0.0089)$, i.e. it is 0.074 ± 0.017 or (0.057, 0.091).

In repeated sampling, approximately 95% of intervals so calculated contain the true excess proportion of colour-blind males relative to the proportion of colour-blind females.

<u>Higher Certificate, Module 3, 2008. Question 3</u>

(i) For the t test for two independent samples, we assume that the underlying populations are Normally distributed with common variance.

The sample sizes are $n_1 = 8$ for library 1 and $n_2 = 9$ for library 2. From the values given, mean borrowing times are 15 for library 1 and 17 for library 2. For library 1, the sample variance is $\{2924 - (120^2/8)\}/7 = 1124/7 = 160.57$; for library 2, it is $\{4359 - (153^2/9)\}/8 = 1758/8 = 219.75$.

[It might be noted at this point that the variance ratio (max/min) is 1.37. Comparison with (say) the upper 5% point of $F_{8,7}$, which is 3.73, shows that the assumption of a common variance is reasonable.]

The pooled estimate of variance is $\dfrac{(7\times 160.57)+(8\times 219.75)}{7+8} = 192.13$.

Thus the value of the test statistic is $\dfrac{17-15}{\sqrt{192.13\left(\frac{1}{8}+\frac{1}{9}\right)}} = 0.297$,

which we refer to t_{15}. This is clearly not significant at any of the usual levels (e.g. the double-tailed 5% point is 2.131), so the null hypothesis is accepted and we conclude that the mean borrowing times at the two libraries may be taken as equal.

(ii) A suitable non-parametric test is the Wilcoxon rank sum test (or, equivalently, the Mann-Whitney version of this test). It tests for difference in the medians of two distributions which are assumed to be otherwise identical. The procedure is to rank the combined samples in ascending order and then add the ranks of the members of the smaller sample.

Value	1	2	3	4	5	6	7	10	12	15	18	21	26	27	36	38	42
Rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Library	2	1	1	2	1	2	2	1	2	1	2	1	1	2	2	1	2

Test statistic = sum of ranks for "library 1" = 2 + 3 + … + 16 = 69. This is referred to the Wilcoxon rank sum test table (Table 10 in the Society's *Statistical tables for use in examinations*), with $n_1 = 8$ and $n_2 = 9$.

A two-sided test is required. We use a 5% significance level. The "0.025" section of the table shows that 51 is just in the lower 2½% tail of the null distribution, so the observed value of 69 is not significant at the lower end of the distribution. To confirm that it is not significant at the upper end, we can either note that it is less than the mean of the distribution (which is given in the Table: $n_1(n_1 + n_2 + 1)/2 = 72$) or use symmetry to argue that 93 is just in the upper 2½% tail. Thus the result is not significant and the null hypothesis is accepted: the medians of the two distributions may be taken as equal.

Solution continued on next page

(iii) For small samples, dotplots are quick to construct and preserve all the information in the data; box and whisker plots would be an acceptable alternative. Dotplots are shown below.

```
Dot plot: Library 1

     .. .    .    .     .    .           .


Dot plot: Library 2

    .  . ..    .     .        .        .     .

   +---------+---------+---------+---------+---------+
   0        10        20        30        40        50
```

The key point is that both samples are clearly positively skew, so that the assumption of underlying Normality, as is required for the *t* test, appears ill-founded. Thus the Wilcoxon test is preferable to the *t* test here. [The fact that the two tests yield similar conclusions may be noted, but this is irrelevant to the fact that the non-parametric test has better validity for these data.]

The dotplots also suggest that the two distributions have comparable spreads. If the *t* test were used, this would be important in terms of the requirement for equal variances, while the Wilcoxon procedure also requires equal "spreads" in that the underlying distributions are supposed to be identical except possibly for different "locations".

<u>Higher Certificate, Module 3, 2008. Question 4</u>

(i) $\Sigma w = 390$, so the sample mean is $\bar{w} = 39$ (grams).

$\Sigma w^2 = 15354$, so the sample variance is $s^2 = \{15354 - (390^2/10)\}/9 = 16$, so that $s = 4$ (grams).

[Alternatively, these values could be obtained directly from calculators.]

(ii) (a) A 95% confidence interval for the true mean weight is based on the t_9 distribution, using the result $\dfrac{\bar{W}-\mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$.

The double-tailed 5% point of this distribution is $t = 2.262$. Hence the interval is given by

$$\bar{w} \pm t\frac{s}{\sqrt{n}} \quad = \quad 39 \pm 2.262 \times \frac{4}{\sqrt{10}},$$

i.e. it is 39 ± 2.86 or (36.14, 41.86) (grams, working to 2 decimal places).

(b) A 95% confidence interval for the variance of the weights is based on the χ_9^2 distribution, using the result $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$.

The lower and upper 2½% points of this distribution are $l = 2.700$ and $u = 19.023$. Hence the interval is given by

$$\left(\frac{(n-1)s^2}{u}, \frac{(n-1)s^2}{l}\right) = \left(\frac{144}{19.023}, \frac{144}{2.700}\right) = (7.57,\ 53.33).$$

Solution continued on next page

(iii) (a) A one-sided test is required. The value of the test statistic is, in the usual notation,

$$\frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}} = \frac{39-42}{\frac{4}{\sqrt{10}}} = -2.37$$

which we refer to t_9. The lower 5% point of this distribution is –1.833, so the result is significant at the 5% level and the null hypothesis that the true mean weight is 42 grams is rejected: there is evidence that it is less than this.

(b) Again a one-sided test is required. The value of the test statistic is, in the usual notation,

$$\frac{(n-1)s^2}{\sigma_0^{\,2}} = \frac{9\times 16}{12} = 12.00$$

which we refer to χ_9^2. The upper 5% point of this distribution is 16.919, so the result is not significant at the 5% level and the null hypothesis that the true variance is 12 (grams squared) is accepted. There is no evidence that it is more than this.