THE ROYAL STATISTICAL SOCIETY

2009 EXAMINATIONS – SOLUTIONS

HIGHER CERTIFICATE

MODULE 3

BASIC STATISTICAL METHODS

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, 2009. Question 1

(i) If the scribe makes errors at random at an average rate that remains constant over time, then the number of errors per page should follow a Poisson distribution.

(ii) The sample mean is $\sum x f_x / 100 = 200/100 = 2$.

(iii) The calculations of the expected frequencies are laid out below. The table of the Poisson cumulative distribution function has been used here. Alternatively, the usual mathematical expression for Poisson probabilities could be used. It is clear that the expected frequency for $x \geq 5$ will be less than 5, so all these cells will be combined in the subsequent calculation of the X^2 statistic.

x	cum freq	difference	expected frequency
0	0.1353		13.53
1	0.4060	0.2707	27.07
2	0.6767	0.2707	27.07
3	0.8571	0.1804	18.04
4	0.9473	0.0902	9.02

The calculations for the usual X^2 statistic are laid out below.

x	O_X	E_X	$(O_X - E_X)^2/E_X$
0	10	13.53	0.9210
1	32	27.07	0.8979
2	24	27.07	0.3482
3	22	18.04	0.8693
4	9	9.02	0.0000
≥5	3	5.27	0.9778
sum			4.0141

We refer the value of X^2, i.e. 4.0141, to χ_4^2.

The upper 5% point of this distribution is 9.488, so the result is not significant at the 5% level and the null hypothesis that the number of errors per page has a Poisson distribution is accepted. Thus there is no evidence to suggest that the number of errors per page do not follow a Poisson distribution, which indicates that it is reasonable to suppose that the scribe makes errors at random.

Higher Certificate, Module 3, 2009. Question 2

(i) A 95% confidence interval for the population mean weight is based on the t_{19} distribution, using the result $\dfrac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$.

The double-tailed 5% point of this distribution is $t = 2.093$. Hence the interval is given by

$$\bar{x} \pm t\frac{s}{\sqrt{n}} \;=\; 99.43 \pm 2.093 \times \frac{\sqrt{0.4678}}{\sqrt{20}},$$

i.e. it is 99.43 ± 0.32 or (99.11, 99.75) (grams).

(ii) A 95% confidence interval for the population variance of the weights is based on the χ^2_{19} distribution, using the result $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$.

The lower and upper 2½% points of this distribution are $l = 8.907$ and $u = 32.852$. Hence the interval is given by

$$\left(\frac{(n-1)s^2}{u}, \frac{(n-1)s^2}{l}\right) \;=\; \left(\frac{8.8882}{32.852}, \frac{8.8882}{8.907}\right) = (0.271,\ 0.988).$$

(iii) We test the null hypothesis H_0: $\mu = 100$ against the alternative hypothesis H_1: $\mu < 100$. So a one-sided test is required. The value of the test statistic is, in the usual notation,

$$\frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}} \;=\; \frac{99.43-100}{\frac{\sqrt{0.4678}}{\sqrt{20}}} \;=\; -3.727$$

which we refer to t_{19}. The lower 1% point of this distribution is –2.539, so the result is significant at the 1% level and the null hypothesis that the true mean weight is 100 grams is rejected. There is strong evidence to reject the manufacturer's claim.

(iv) The p-value of a test statistic is the probability under the null hypothesis of obtaining the calculated test statistic or a more extreme value. In the present case, from the table of the t distribution, the test statistic is significant at the 0.001 level, but not at the 0.0005 level. Hence $0.0005 < p < 0.001$.

Higher Certificate, Module 3, 2009. Question 3

(i) We have a 2×2 contingency table. The null hypothesis is that there is no association between district and uptake of digital television. The contingency table is as follows, with the expected frequencies in brackets in each cell (e.g. 32.5 = 65 × 50 / 100).

		District A	*District B*	Total
Digital TV	*Yes*	38 (32.5)	27 (32.5)	65
	No	12 (17.5)	23 (17.5)	35
	Total	50	50	100

All the differences between observed and expected frequencies are ±5.5, becoming ±5.0 if Yates' correction is used. Thus the usual test statistic can be calculated as (using Yates' correction)

$$(5.0)^2\left\{\frac{1}{32.5}+\frac{1}{32.5}+\frac{1}{17.5}+\frac{1}{17.5}\right\} = 4.396$$

(or 5.319 if Yates' correction is not used). This is referred to χ_1^2.

The 5% point of this distribution is 3.841, the 1% point is 6.635. So the result is significant at the 5% level but not at the 1% level. There is some evidence that there is an association between district and uptake of digital television. The data indicate that uptake is greater in district A, so we have evidence that this is a real effect, i.e. there is evidence of a greater uptake in district A.

(ii) p_B, the proportion of households in district B which have digital television, is estimated by $\hat{p}_B = \frac{27}{50} = 0.54$. The estimated variance of $\hat{p}_B$ is given by

$$\frac{\hat{p}_B\left(1-\hat{p}_B\right)}{n_B} = \frac{0.54\times 0.46}{50} = 0.004968.$$

Thus the approximate 95% confidence interval for p_B is given by 0.54 ± (1.96×√0.004968), i.e. it is (0.402, 0.678).

(iii) McNemar's test should be used. The test statistic is given by

$$(15-9)^2/(15+9) = 1.5$$

[If the continuity correction is used, the test statistic is $(15-9-1)^2/(15+9) = 1.042$.]

This is referred to χ_1^2. The 5% point of this distribution is 3.841. So the result is not significant, even at the 5% level. There is no evidence of any difference between the levels of uptake of broadband in the two districts.

Higher Certificate, Module 3, 2009. Question 4

(i) The appropriate parametric test is the paired comparison t test. It is assumed that the differences d_i between the marks assigned by the two examiners are a random sample from a Normal distribution.

From the data, we have that the sample mean for the differences is $\bar{d} = 6.42$ and the sample variance is $s_d^2 = 187.90$. Thus the test statistic for testing that the population mean marks for the two examiners are the same, against the (two-sided) alternative that they differ, is

$$\frac{\bar{d}(-0)}{s_d\sqrt{\frac{1}{12}}} = \frac{6.42}{3.957} = 1.62,$$

which is referred to t_{11}. This is not significant at the 5% level (double-tailed 5% point is 2.201), so there is no evidence to reject the null hypothesis – it seems that the population mean marks for the two examiners are the same.

(ii) The Wilcoxon signed rank test is suitable (the sign test could also be used here). The procedure is to rank the absolute values of the differences, then re-attach the signs to the ranks, find the rank sum for the ranks with negative signs and for the ranks with positive signs, and consider the smaller of these (it is usually convenient to choose the smaller because *lower* critical points are given in the tables). The necessary calculations are shown below.

Student	1	2	3	4	5	6	7	8	9	10	11	12
Difference d	–11	–10	3	9	–7	2	–6	8	21	28	14	26
Rank of $\lvert d \rvert$	8	7	2	6	4	1	3	5	10	12	9	11

Sum of ranks with positive signs = 56.

Sum of ranks with negative signs = 22.

The smaller rank sum is 22. For a two-sided test, say at the 5% level, we refer this to the lower 2½% point for the W_{12} distribution as shown in the Society's statistical tables for use in examinations. This is 13 so, at the 5% level of significance, we cannot reject the null hypothesis. It appears that the two examiners assign the same marks on the whole.

[Note. If the sign test is used, examiner A gives the higher mark 4 times out of 12, leading to a non-significant result at the 5% level.]

(iii) The Wilcoxon signed rank (or sign) test is more robust in that it does not require the Normality assumption. On the other hand, if the assumption of Normality for the population of differences is justified, the t test will be more powerful.

In the present case, we arrive at the same conclusion whichever of the tests we use.