THE ROYAL STATISTICAL SOCIETY

2005 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, 2005. Question 1

(i) The Poisson distribution to explain numbers of goals might be a reasonable assumption if home team scores can be regarded as random events occurring at a constant average rate throughout the season. If so, the number of home team goals in a match is Poisson with parameter (mean) equal to this constant average rate, μ say.

(ii) $\bar{r} = 634/380 = 1.6684$.

$$s^2 = \frac{1}{\Sigma f - 1}\left\{\Sigma fr^2 - \frac{(\Sigma fr)^2}{\Sigma f}\right\} = \frac{1}{379}\left(1778 - \frac{634^2}{380}\right) = 1.9003.$$

(iii) We take μ as 1.6684. So $P(R=0) = e^{-1.6684} = 0.1885$, and the expected frequency for r = 0 is 380 × 0.1885 = 71.65.

Similarly, $P(R=1) = 1.6684e^{-1.6684} = 0.3145$, and the expected frequency for r = 1 is 119.51.

Hence we have (taking the remaining expected frequencies from the question paper)

r	0	1	2	3	4	≥5	Total
Observed	81	112	101	44	28	14	380
Expected	71.65	119.51	99.72	55.46	23.13	10.51	379.98

[Note. There is a very small rounding error in the calculations of expected frequencies.]

The test statistic is

$$X^2 = \sum \frac{(O-E)^2}{E} = \frac{(81-71.65)^2}{71.65} + \frac{(112-119.51)^2}{119.51} + \ldots + \frac{(14-10.51)^2}{10.51} = 6.261,$$

which is referred to χ_4^2 (note 4 degrees of freedom because the table has 6 cells and there is one estimated parameter). This is not significant (the 5% point is 9.49); we cannot reject the null hypothesis, i.e. there is no evidence against the Poisson model with these data.

For the test, the expected frequencies need to be not too small (≥5 is often used as a criterion). This would not be the case if frequencies for large r were not combined.

(iv) The negative binomial is commonly used where there is "over-dispersion". [It assumes that the rate (μ) is not always constant but varies (from match to match) according to a gamma distribution.]

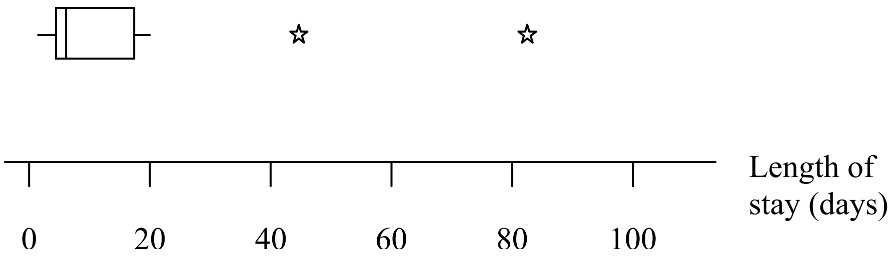
Higher Certificate, Paper II, 2005. Question 2

(i) The ranked data are as follows, with the median and the lower and upper quartiles underlined. [Note. Some slightly different definitions of quartiles are also in use. These would make, at most, only a small difference here.]

1, 1, 2, 3, 3, $\underline{4}$, 4, 5, 5, 6, 6, $\underline{6}$, 7, 7, 8, 9, 14, $\underline{17}$, 19, 19, 20, 44, 82.

Q_1 M Q_3

We have $1.5(Q_3 - Q_1) = 1.5(17 - 4) = 19.5$. So 44 and 82 may be "outliers". These are indicated by stars in the box and whisker plot.



Even apart from the two outliers, the distribution is very skew to the right. This may be explained by some of the admissions being in serious enough condition to need extra care.

(ii) Here a *t* test would be used to examine the hypothesis about the mean duration of bed occupancy. It relies on the data being a sample from a Normal distribution, at least approximately, but this assumption is clearly not valid here. More generally, because of the skewness of the underlying distribution, inferences based on the mean (and standard deviation) of a sample will be unreliable unless a very large sample is available. Any statistical test based on the mean of a small sample will be worthless. Even the sample of size 100 in part (iii) is not really "large" for a case so skew as this.

The solution to part (iii) is on the next page

(iii) We regard the sample of size 100 as being "large" and invoke the Central Limit Theorem so as to use a test based on N(0, 1) [a test based on t_{99} could also be reasonably justified; t_{99} is very close to N(0, 1)].

We have $\bar{x} = 14.88$ and $s^2 = \dfrac{1}{99}\left(44632 - \dfrac{1488^2}{100}\right) = 227.1774$.

The null hypothesis is that $\mu = 14$, the alternative hypothesis is $\mu > 14$, where μ is the (population) mean duration of bed occupation.

The test statistic is $\dfrac{14.88 - 14}{\sqrt{\dfrac{227.1774}{100}}} = \dfrac{0.88}{1.507} = 0.584$, which is clearly not significant as an observation from N(0, 1). There is no evidence that the mean duration is greater than 14 days.

As discussed in parts (i) and (ii), it is clear from the original 23 items of data that the underlying distribution is very skew. Even with a sample of size 100, the result of the test should not be taken as very reliable. (Another illustration of this is provided by calculating a (say) 95% confidence interval in the usual way: $\bar{x} \pm 1.96 \times 1.507$ gives the interval (11.93, 17.83), which is wide for a sample of this size, indicating imprecise results.) The real problem is that the mean does not give useful information about the "typical" length of stay. There is a wider question as to whether "performance" is validly measured by length of stay in any case.

Higher Certificate, Paper II, 2005. Question 3

(i) The expected frequencies on the null hypothesis of no difference between the sexes in the response are found in the usual way from the marginal totals (e.g. that for "Female, No" is 29×35/50 = 20.3). Thus the observed and expected frequencies are

	Observed frequencies				**Expected frequencies**	
	Female	*Male*	**Total**		*Female*	*Male*
No	18	17	35		20.3	14.7
Yes	11	4	15		8.7	6.3
Total	29	21	50			

All the differences between observed and expected frequencies are ±2.3, becoming ±1.8 if Yates' correction is used. Thus the usual test statistic can be calculated as (using Yates' correction)

$$(1.8)^2\left\{\frac{1}{20.3}+\frac{1}{8.7}+\frac{1}{14.7}+\frac{1}{6.3}\right\} = 1.267$$

(or 2.07 if Yates' correction is not used). This is referred to χ_1^2; the upper 5% point is 3.84, so we have no evidence of a real sex difference.

(ii) $p_f - p_m$ is estimated by $\hat{p}_f - \hat{p}_m = \frac{11}{29} - \frac{4}{21} = 0.3793 - 0.1905 = 0.1888$. The estimated variance of $\hat{p}_f - \hat{p}_m$ is given by

$$\frac{\hat{p}_f\left(1-\hat{p}_f\right)}{n_f}+\frac{\hat{p}_m\left(1-\hat{p}_m\right)}{n_m} = 0.0081184+0.0073426 = 0.015461.$$

Thus the approximate 95% confidence interval is given by 0.1888 ± (1.96×√0.015461) i.e. it is (–0.0548, 0.4324) or, in percentage terms, (–5.48%, 43.24%).

The Normal approximation is unlikely to be very good with these small samples, especially as the values of $\hat{p}_f$ and $\hat{p}_m$ suggest that p_f and p_m are some way from 0.5.

(We might note also that the confidence interval is very wide; it does not give much information, due to lack of sufficient data.)

Higher Certificate, Paper II, 2005. Question 4

(i) McNemar's test is required because the samples are paired.

Denoting the entries in the table by $\begin{matrix} a & b \\ c & d \end{matrix}$, the test statistic for McNemar's test is $\dfrac{(|b-c|-1)^2}{b+c}$, with approximate null distribution χ_1^2, the null hypothesis here being that there is no difference between the proportions (probabilities) for the MAT and ELISA tests. (Notice that McNemar's test uses the information from the "discordant" cells of the table.)

Thus the test statistic is $\dfrac{(|25-41|-1)^2}{25+41} = \dfrac{15^2}{66} = 3.409$. This is referred to χ_1^2; the upper 5% point is 3.84, so there is insufficient evidence to say that there is a real difference.

(ii) Approximate 95% confidence intervals for the proportion of positive test results given by each test use the whole data. For MAT, $\hat{p}_M = \frac{92}{462} = 0.1991$; for ELISA, $\hat{p}_E = \frac{108}{462} = 0.2338$.

(a) The estimated variance of $\hat{p}_M$ is (0.1991)(0.8009)/462 = 0.000345135, so the estimated standard deviation is 0.0186. Thus a 95% confidence interval for p_M is given by, approximately, 0.1991 ± (1.96)(0.0186), i.e. it is (0.163, 0.236).

(b) The estimated variance of $\hat{p}_E$ is (0.2338)(0.7662)/462 = 0.000387744, so the estimated standard deviation is 0.0197 Thus a 95% confidence interval for p_E is given by, approximately, 0.2338 ± (1.96)(0.0197), i.e. it is (0.195, 0.272).

Neither of these intervals contains the proposed value of 0.069 – in fact, the intervals are a considerable distance away from that. So neither is consistent with this value.

Higher Certificate, Paper II, 2005. Question 5

(i) The two samples should be from independent Normal distributions with the same variance but possibly different means (the null hypothesis is usually that the two means are equal). The samples are random samples and are independent of each other.

(ii)

$n_1 = 7;\ \ \bar{x}_1 = 54.56,\ s_1^{\ 2} = 534.6956.\qquad n_2 = 14;\ \ \bar{x}_2 = 49.29,\ s_2^{\ 2} = 261.3001.$

The "pooled estimate" of variance is $s^2 = \dfrac{(n_1 - 1)s_1^{\ 2} + (n_2 - 1)s_2^{\ 2}}{n_1 - 1 + n_2 - 1} = 347.6355.$

The test statistic for testing the null hypothesis $\mu_1 - \mu_2 = 0$, where μ_1 and μ_2 are the respective population mean prices, is

$$\frac{\bar{x}_1 - \bar{x}_2 - 0}{s\sqrt{\frac{1}{7} + \frac{1}{14}}} = \frac{5.27}{8.631} = 0.611,$$

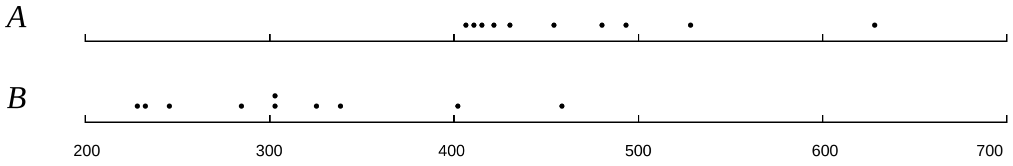
which is referred to t_{19}. This is not significant, so the null hypothesis cannot be rejected. There is no evidence that the true means in the two towns differ.

(iii) Town 1 has greater population and greater numbers of all types of properties, yet only half the sample size was used compared with town 2. The probabilities of selection in the two towns are thus very different. Also, the samples were restricted to the "Property Supplement" and to agents dealing in both towns. The assumption of randomness is doubtful, even whether we have representative samples. The test based on these data must be suspect for practical reasons, even if Normality and constant variance are acceptable. [The usual $F_{6,13}$ test for equality of population variances gives a test statistic value of 2.05 which is not significant.]

Higher Certificate, Paper II, 2005. Question 6

(i) The Mann-Whitney *U* test is preferred to the *t* test for comparing location in two independent samples with the same underlying dispersion if the data come from distributions that are not (approximately) Normal and if the data are ranked rather than measured exactly (i.e. the data are ordinal but not of interval type).

(ii)



Both distributions are skew to the right, of fairly similar shape. The ranges are about the same, suggesting that the underlying dispersions might reasonably be taken as equal. The locations are clearly different. The samples are certainly to small for the Central Limit Theorem to apply to their means.

(iii) The Mann-Whitney *U* test (equivalently, a Wilcoxon rank sum test could be used) is applied as follows. The data and ranks are shown in the table, using average ranks for ties.

231	233	249	285	301	301	328	343	400	407
1	2	3	4	5½	5½	7	8	9	10
B	B	B	B	B	B	B	B	B	A
410	416	421	432	456	460	481	491	532	634
11	12	13	14	15	16	17	18	19	20
A	A	A	A	A	B	A	A	A	A

$n_1 = 10$, $n_2 = 10$. Total rank for component type *A* is $T_A = 149$; for *B* is $T_B = 61$.

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

$$U_1 = n_1 n_2 + \tfrac{1}{2} n_1 (n_1 + 1) - T_A = 100 + 55 - 149 = 6.$$

$$U_2 = n_1 n_2 + \tfrac{1}{2} n_2 (n_2 + 1) - T_B = 100 + 55 - 61 = 94.$$

So $U_{min} = 6$. From tables, the critical value for a *U* test with $n_1 = n_2 = 10$ at the 5% two-tailed level is 23. As 6 < 23, we reject the null hypothesis at the 5% level of significance. In fact we would also reject at the 1% level. So (in a form for the non-statistician to understand) there is extremely strong evidence that the lifetimes of the two types of components are different and we can strongly conclude that, on the whole, lifetimes of type *A* are longer than those of type *B*.

Higher Certificate, Paper II, 2005. Question 7

(i) $y_{ij} = \mu + t_i + \varepsilon_{ij}, \quad i = 1, 2, ..., k, \quad j = 1, 2, ..., r_i, \quad \{\varepsilon_{ij}\} \sim \text{ind N}(0, \sigma^2).$

There are k treatments, indexed by i = 1, 2, …, k. In the experiment or survey, there are r_i units (individuals) in the ith group, i.e. receiving the ith treatment. y_{ij} is the observation (response) for the jth individual in group i.

μ is the overall population general mean. t_i is the population mean effect (departure from μ) due to treatment i, with $\Sigma_i t_i = 0$.

The Normally distributed residual (error) terms ε_{ij} all have variance σ^2 and are uncorrelated (independent).

This is an additive model: the components add together, and together explain all the variation in the responses.

(ii) The "treatments" here are "high", "low" and "work". $r_1 = r_2 = r_3 = 12$.

Totals are:	High	Low	Work
	5528	3754	3511

The grand total is 12793. $\Sigma\Sigma y_{ij}^2 = 5719139$.

"Correction factor" is $\dfrac{12793^2}{36} = 4546134.694$.

Therefore total SS = 5719139 – 4546134.694 = 1173004.306.

SS for treatments = $\dfrac{5528^2}{12} + \dfrac{3754^2}{12} + \dfrac{3511^2}{12} - 4546134.694 \ = \ 202067.056$.

The residual SS is obtained by subtraction.

Hence the analysis of variance table is as follows (SS and MS entries are slightly rounded).

SOURCE	DF	SS	MS	*F* value
Treatments	2	202067	101034	3.43 Compare $F_{2,33}$
Residual	33	970937	29422	$= \hat{\sigma}^2$
TOTAL	35	1173004		

The upper 5% point of $F_{2,33}$ is about 3.3; the treatments effect is significant. There is evidence to reject the null hypothesis that all treatments have the same effect.

Solution continued on the next page

To investigate treatment differences, first calculate the treatment means, which are (in ascending order, for clarity)

$$\text{Work}: \ 292.58 \qquad \text{Low}: \ 312.83 \qquad \text{High}: \ 460.67$$

The least significant difference between any pair of these means is

$$t_{33}\sqrt{\frac{2\times 29422}{12}} = 70.026\,t_{33} \qquad \text{where } \ t_{33} = \begin{cases} 2.035 & \text{at } 5\% \\ 2.736 & \text{at } 1\% \\ 3.617 & \text{at } 0.1\% \end{cases}$$

so the least significant differences are 142.50 for 5%, 191.59 for 1% and 253.28 for 0.1%. Thus the only apparent difference is that "high" gives a larger mean response than "low" and "work", judged at the 5% level; "low" and "work" do not differ.

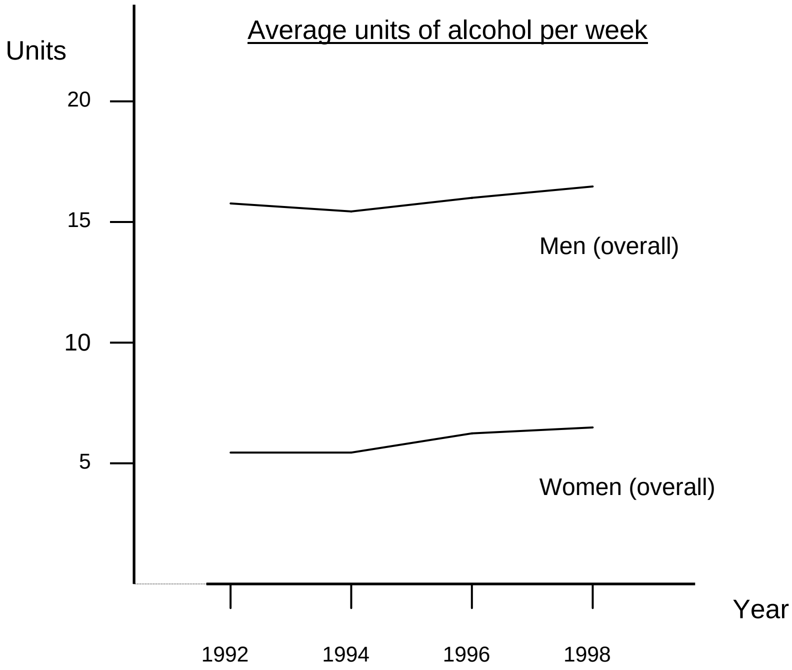
Report

After carrying out an analysis which compares group means against internal variability of responses in the groups, we find some evidence that "high" shows more persistence than the other two groups, whose results are quite similar. The within-group variability is very high.

Higher Certificate, Paper II, 2005. Question 8

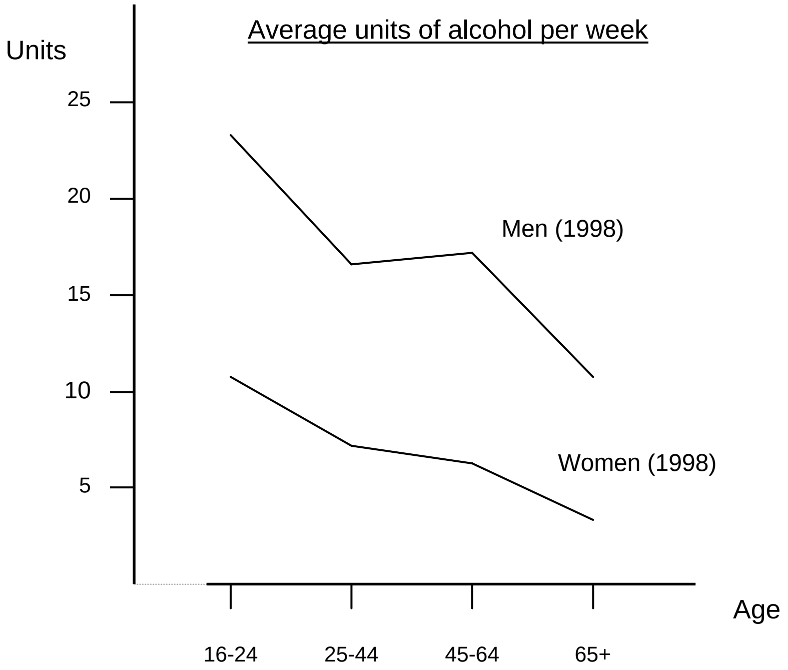
Part (i)

(a)



(The limits of electronic reproduction may make the lines in this diagram and the next appear somewhat ragged.)

(b)



The solution to part (ii) is on the next page

Part (ii)

For overall consumption of alcohol, there was a slight increase over the time period as a whole, but not a very regular pattern.

Both for men and for women, the 16–24 age groups showed a distinct rise between the first two and the last two data sets (i.e. between 1992/94 and 1996/98).

Both for men and for women, the 65 and over age group showed a fall over the last two periods (i.e. from 1996 to 1998). This was also true of the 25–44 age groups, but the other age groups showed quite substantial increases.

Overall, there is a general decrease in consumption with age, though this is largely explained by markedly high 16–24 and low 65+ figures.

Overall, men drink two to three times as much as women.