THE ROYAL STATISTICAL SOCIETY

2001 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, 2001. Question 1

(i) If random samples are taken from a non-Normal population, whose mean and standard deviation are known to be μ and σ, then when the sample size n is large the sample mean ($\bar{X}$, say) will be approximately Normal with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$. The approximation improves as n increases, and is adequate for moderate size n if the X distribution is not very skew. The total of n observations has a similar distribution (scaled up by a factor n). Thus estimators that are averages or totals can often be taken as approximately Normal.

(ii) Let N, S be the new and standard sprays respectively; then the mean $\bar{X}_N - \bar{X}_S$ can be taken as N(249 − 237, $\frac{490}{100}+\frac{410}{100}$) approximately (inserting the sample estimates for the mean and variance), i.e. N(12,9).

A 95% confidence interval then is $12 \pm (1.96)\sqrt{9} = 12 \pm 5.88$

or 6.12 to 17.88 kg for $\mu_N - \mu_S$.

The unbiased estimate of $(\mu_N - \mu_S)$ is 12 kg, and the interval (6.12, 17.88) contains the true value of $(\mu_N - \mu_S)$ with probability 0.95. There is strong evidence to say that N is better than S.

(iii) $\hat{p}_N = \frac{40}{50} = 0.8$, the proportion of healthy plants on N; and $\hat{p}_S = \frac{70}{100} = 0.7$ on S. Also $n_N = 50$ and $n_S = 100$. Using a Normal approximation to the binomial distribution, the true difference $p_N - p_S \sim \mathrm{N}\left(\hat{p}_N - \hat{p}_S, \frac{\hat{p}_N(1-\hat{p}_N)}{50} + \frac{\hat{p}_S(1-\hat{p}_S)}{100}\right)$

A 95% confidence interval for $p_N - p_S$ is $(0.8 - 0.7) \pm 1.96\sqrt{\frac{0.8 \times 0.2}{50} + \frac{0.7 \times 0.3}{100}}$

i.e. $0.1 \pm 1.96 \times 0.0728$ or 0.1 ± 0.143, i.e. (−0.04, 0.24).

It is possible that S may be better by 4%, but the upper limit is for N to be better by 24%.

<u>Higher Certificate, Paper II, 2001. Question 2</u>

(i) If all dice are thrown "fairly" (and all are "fair" in construction), and throws are independent, then the conditions for a binomial distribution are satisfied; p = 1/6 since there are six equally possible results, and n = 5, the "sample size" (number thrown) each time. The number of "success" (sixes) is counted (R).

(ii) $$P(R=r)=\binom{5}{r}\left(\frac{1}{6}\right)^r\left(\frac{5}{6}\right)^{5-r}$$ for r = 0, 1, 2, 3, 4, 5.

Expected frequencies are these probabilities $\times$ 200.

$$P(0)=\left(\frac{5}{6}\right)^5=0.40188,\quad P(1)=5\left(\frac{1}{6}\right)\left(\frac{5}{6}\right)^4=0.40188,\quad P(2)=10\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^3=0.16075$$

$$P(3)=10\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^2=0.03215,\quad P(4)=5\left(\frac{1}{6}\right)^4\left(\frac{5}{6}\right)=0.00322,\quad P(5)=\left(\frac{1}{6}\right)^5=0.00013$$

Compare these <u>expected</u> values with those <u>observed</u>, in the form of frequencies, using a χ^2 test. Group 3, 4, 5 together to prevent very small expected frequencies.

r	0	1	2	3	4	5	Total
Observed	90	78	26	4	1	1	200
				combine	these to give 6		
Expected	80.38	80.38	32.15		7.10		(200.01)

$$X^2=\frac{(90-80.38)^2}{80.38}+\frac{(78-80.38)^2}{80.38}+\frac{(26-32.15)^2}{32.15}+\frac{(6-7.10)^2}{7.10}=2.569$$ which is

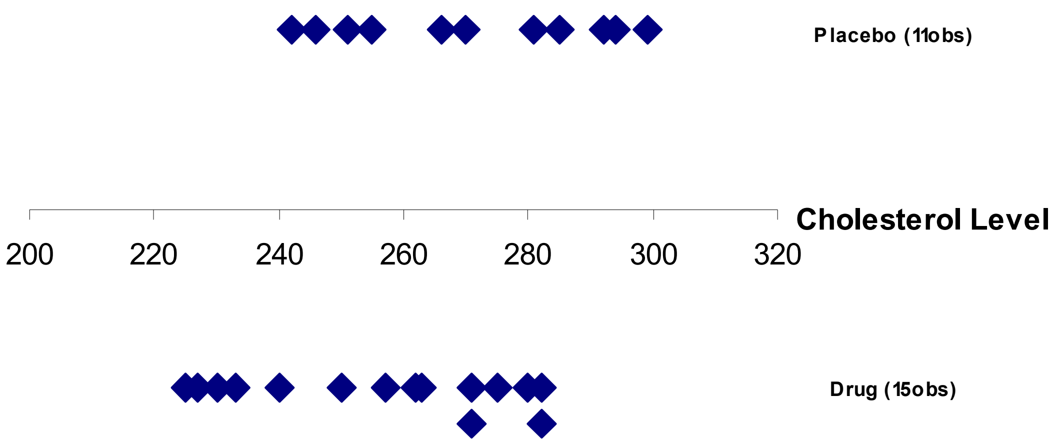
not significant as an observation from χ^2_3 (3 degrees of freedom since we are <u>given</u> the value of p).

The null hypothesis is that R is binomial with n = 5, p = 1/6, and there is no evidence that the observed frequencies depart seriously from those expected on this hypothesis.

We may assume the binomial model explains the data, and therefore the conditions for a binomial do apply.

Higher Certificate, Paper II, 2001. Question 3

(i) Both data sets fairly symmetrical, but not clustered round mean.



Placebo on average somewhat higher than Drug.

(ii)

Rank	1	2	3	4	5	6	7	8	9	10	11	12	13
Obs.	225	227	230	233	240	242	246	250	251	255	257	262	263
Trt.	D	D	D	D	D	P	P	D	P	P	D	D	D

14	15	16	17	18	19	20	21	22	23	24	25	26
266	270	271	271	275	280	281	282	282	285	292	294	299
P	P	D	D	D	D	P	D	D	P	P	P	P

A Mann-Whitney U test may be applied.

Sum of ranks of P = 179. $U_P = (11\times 15) + \frac{1}{2}(11\times 12) - 179$

$= 52$

The 5% one-sided critical value is 44 for n_1=11, n_2=15.

Therefore on these data there is no evidence for claiming that the drug reduces blood pressure.

(iii) If we assume the data to be Normally distributed, with the same σ^2 in each distribution, a t test can be applied.

Placebo: $\bar{x} = 271.00,\ s^2 = (20.489)^2$

Drug: $\bar{x} = 256.53,\ s^2 = (20.908)^2$

clearly $s^2{}_{\mathrm{P}}$, $s^2{}_{\mathrm{D}}$ can be pooled to give:

$$s^2 = \frac{(10 \times 20.489^2) + (14 \times 20.908^2)}{24} = 429.9172 = (20.734)^2 .$$

$$\mathrm{H}_0 : \mu_{\mathrm{D}} = \mu_{\mathrm{P}}, \quad \mathrm{H}_1 : \mu_{\mathrm{D}} < \mu_{\mathrm{P}}$$

$$t_{24} = \frac{\bar{x}_{\mathrm{P}} - \bar{x}_{\mathrm{D}}}{SE(\bar{x}_{\mathrm{P}} - \bar{x}_{\mathrm{D}})} = \frac{271.00 - 256.53}{20.734\sqrt{\frac{1}{11} + \frac{1}{15}}} = \frac{14.47}{8.23} = 1.758$$

which is greater than the one-tail 5% value which is 1.711.

Hence there is evidence to claim a reduction using the drug.

(iv) The assumption of Normality increases the power of the t test compared with Mann-Whitney which makes no distributional assumption.

Higher Certificate, Paper II, 2001. Question 4

(i) We need to assume that the lifetimes follow a Normal distribution, i.e. the data are independent, identically distributed with variance σ^2.

Then $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$.

The estimated variance $s^2 = 401.143$ with 7 d.f.

The null hypothesis is $\sigma^2 = 625$, so the test statistic is $\frac{7 \times 401.143}{625} = 4.493$. Comparing with χ^2_7, this is not significant, so the null hypothesis (strictly $\sigma^2 \leq 625$) is not rejected.

(ii) The null hypothesis is $\sigma_1^2 = \sigma_2^2$ where σ_i^2 is the variance on process *i*, and the alternative hypothesis will be $\sigma_1^2 > \sigma_2^2$. (Again the null hypothesis is, strictly, $\sigma_1^2 \leq \sigma_2^2$.)

The ratio $\frac{S_1^2}{S_2^2} \sim F_{n_1-1,\, n_2-1}$. Estimates each have 9 d.f. $s_1^2 = 384.667$, $s_2^2 = 166.622$.

Test statistic is $\frac{384.667}{166.622} = 2.309$, not significant on $F_{9,9}$ since the one-tail 5% point is 3.18.

Although there is a reduction, on this evidence it is not significant and the null hypothesis cannot therefore be rejected. We must assume the variance has not been reduced. Again the Normality of both sets of data must be assumed.

Higher Certificate, Paper II, 2001. Question 5

(a) A suitable null hypothesis is that each judge is equally likely to choose either; hence N, the number preferred, is binomial with parameters 12 and $p = \frac{1}{2}$. We have $n_A = 4$ and $n_B = 8$.

$$P(N \le 4) \text{ in } B\left(12, \frac{1}{2}\right) = \sum_{r=0}^{4} \frac{12}{r!(12-r)!}\left(\frac{1}{2}\right)^{12} = \frac{1}{2^{12}}\{1+12+66+220+495\}$$

$$= \frac{794}{4096} = 0.194 .$$

This result is not "unlikely", and the null hypothesis cannot be rejected. We have no conclusive evidence to say which may be more popular.

(b) (i) McNemar's test deals with paired data, as these are.

H_0: there is no association between stress and success,
H_1: there is such an association.

The test uses the two terms in the right-to-left diagonal, giving test statistic

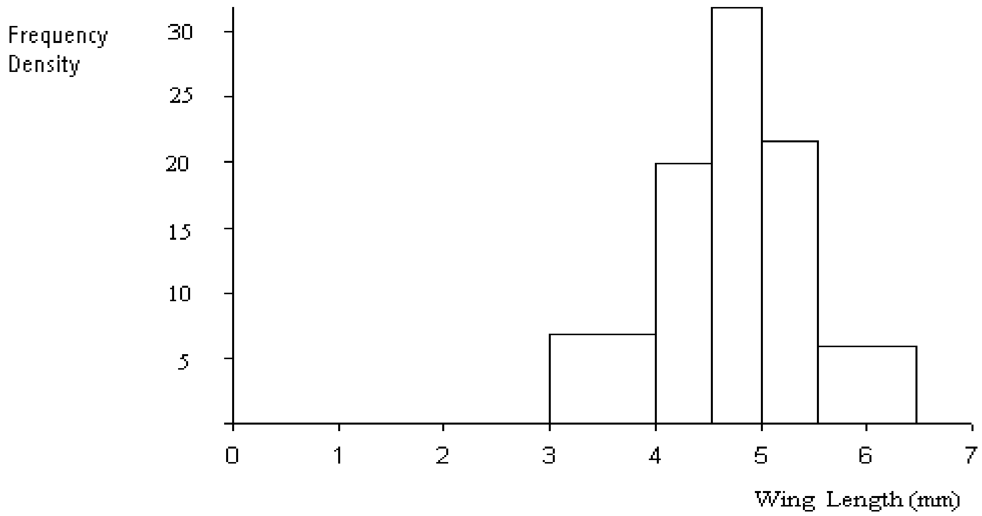
$$\frac{(9-20)^2}{9+20} = 4.172$$

Comparing with χ_1^2, this is significant at the 5% level. So there is evidence against the null hypothesis, in favour of the existence of some association.

(ii) The standard 2×2 test has given a non-significant result $\left(\chi_1^2 < 3.84\right)$. It does not use the matched nature of the data, so it has less power than McNemar's test to look for association. In this example, McNemar's result is based on the proportions of successful individuals in the two samples "stress" and "no stress".

Higher Certificate, Paper II, 2001. Question 6

(i)



Interval	Midpoint, x	f	fx	fx^2	F
3.0 – 4.0	3.50	14	49.0	171.500	14
4.0 – 4.5	4.25	20	85.0	361.250	34
4.5 – 5.0	4.75	32	152.0	722.000	66
5.0 – 5.5	5.25	22	115.5	606.375	88
5.5 – 6.5	6.00	12	72.0	432.000	100
		100	473.5	2293.125	

$\bar{x} = 4.735$. $s^2 = \frac{1}{99}\left(2293.125 - \frac{473.5^2}{100}\right) = 0.5162$; $s = 0.718$.

Median $\approx 4.5 + \frac{16}{32} \times 0.5 = 4.75$ (approx – depending on accuracy of measurement).

The data appear to be approximately symmetrical $(\bar{x} \approx \text{median})$ and, from the histogram, could be assumed Normally distributed.

(ii) With these assumptions, and based on the given data, a 95% confidence interval for the true mean is $\bar{x} \pm 1.96\frac{s}{\sqrt{n}}$

i.e. $4.375 \pm 1.96 \times 0.0718$, or 4.735 ± 0.141, i.e. $(4.59 \text{ to } 4.88)$.

<u>Higher Certificate, Paper II, 2001. Question 7</u>

(i) $y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$, where y_{ij} is the observation on the jth unit receiving treatment i (or lying in group i), μ is a grand mean, α_i is an 'effect' (departure from μ) due to treatment (or group) i and ε_{ij} is a random, Normally distributed, residual 'error' term, all $\{\varepsilon_{ij}\}$ independent and all having variance σ^2.

The model is "additive", constructed by adding quantities rather than (for example) multiplying them; this and the properties of $\{\varepsilon_{ij}\}$ are necessary assumptions for the analysis.

(ii)

Woodland	n	Σx	Σx^2	$\bar{x}$	$(\Sigma x)^2/n$
A	10	664	45780	66.400	44089.600
B	8	313	13453	39.125	12246.125
C	6	402	27938	67.000	26934.000
	N=24	1379=G	87171		83269.725

$$\frac{G^2}{N} = 79235.042$$

$$\text{Total corrected SS} = 87171 - \frac{G^2}{N} = 7935.958$$

$$\text{Woodlands SS} = 83269.725 - \frac{G^2}{N} = 4034.683$$

<u>Analysis of Variance</u>

ITEM	DF	SUM OF SQUARES	MEAN SQUARE	
Woodlands	2	4034.683	2017.342	$F_{2,21} = 10.86$
Residual	21	3901.275	185.775=s^2	
TOTAL	23	7935.958		

Comparing 10.86 with $F_{2,21}$, the null hypothesis "all α_i are zero" can be rejected at the 0.1% significance level.

From the values of $\bar{x}$, we can immediately see that the reason for this is that B is different from the other two.

Compare B and C:

$\bar{x}_C - \bar{x}_B = 27.875$, SE of difference $= \sqrt{s^2\left(\frac{1}{8}+\frac{1}{6}\right)}$.

Hence the t_{21} test statistic is $\frac{27.875}{7.361} = 3.79$ which is significant at the 0.1% level.

Alternatively, a 95% confidence interval for $(\mu_C - \mu_B)$ is

$27.875 \pm 2.080 \times 7.361$ i.e. $(12.56 \text{ to } 43.19)$.

The precision of the results is poor, as shown by the wide interval.

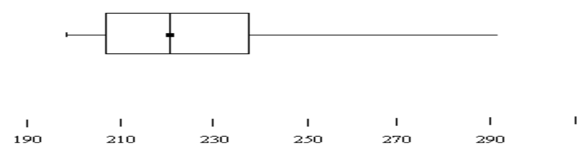
There is a similar result for B versus A.

Higher Certificate, Paper II, 2001. Question 8

(i) Total marks ranked in order of size from lowest:

$$197, 200, 203, \underset{q}{\underline{\underline{206}}}, 210, 213, 217, \underset{M}{\underline{\underline{220}}}, 223, 234, 237, \underset{Q}{\underline{\underline{238}}}, 245, 273, 290$$

Median = 220. Lower quartile = 206 (or, with an alternative definition, ½(206 + 210) = 208); upper quartile 238 (or 237 ½).



An alternative display marks the whisker ending at 273, with 290 shown as *. The distribution is skew, since the median is not in the centre of the box and there is a very long whisker to the right – although this is largely caused by the top two observations.

(ii) To compare the rank orders of Practical and Written, the 15 students' marks need to be ranked and the difference in ranks, d, found. Then Spearman's coefficient is $1-\dfrac{6\sum d^2}{n\left(n^2-1\right)}$.

Student	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
P	4	12	2	11	6	10	8	1	9	13	15	5	14	3	7
W	9	10	2	13	12	11	8	6	7	4	15	5	14	3	1
d	–5	2	0	–2	–6	–1	0	–5	2	9	0	0	0	0	6

$\sum d^2 = 216$. $r_S = 1-\dfrac{6\times 216}{15\times 224} = 0.614$.

Both this and the product-moment coefficient are significant at the 1% level, so there is firm evidence that the practical and written marks increase together. The ranking pattern is disturbed by (10), so reducing r_S. The diagram shows a basically linear relation with some noticeable scatter.