EXAMINATIONS OF THE ROYAL STATISTICAL SOCIETY

(formerly the Examinations of the Institute of Statisticians)

HIGHER CERTIFICATE IN STATISTICS, 2001

Paper III : Statistical Applications and Practice

Time Allowed: Three Hours

Candidates should answer **FIVE** *questions.*

All questions carry equal marks.
The number of marks allotted for each part-question is shown in brackets.

Graph paper and Official tables are provided.

Candidates may use silent, cordless, non-programmable electronic calculators.

Where a calculator is used the **method** *of calculation should be stated in full.*

Note that $\begin{pmatrix} n \\ r \end{pmatrix}$ *is the same as* ${}^nC_r$ *and that* ln *stands for* $\log_e$.

This examination paper consists of 10 printed pages. This front cover is page 1. The reverse of the front cover, which is intentionally left blank, is page 2. Question 1 starts on page 3.

There are 8 questions altogether in the paper.

©RSS 2001

1. The following data are blood cholesterol levels (in mg/100ml) of 10 heart attack patients one and two weeks after the attack, together with the difference between the two levels.

Patient	*One week after*	*Two weeks after*	*Difference*	*Total*
1	142	116	–26	258
2	360	352	–8	712
3	244	269	+25	513
4	186	190	+4	376
5	210	215	+5	425
6	236	242	+6	478
7	288	248	–40	536
8	276	220	–56	496
9	224	200	–24	424
10	311	302	–9	613
Total	2477	2354	–123	4831

(i) Perform a paired *t* test on these data.

(3)

(ii) Calculate the sums of squares (SSs) for "patients" and "week" and perform the analysis of variance (ANOVA) for a randomised block design, with "patients" corresponding to blocks. Note that the (corrected) total SS = 74498.95.

(6)

(iii) Perform a Wilcoxon signed-rank test on these data.

(4)

(iv) State the null and alternative hypotheses being tested in parts (i), (ii) and (iii) and the assumptions the tests involved make. What conclusions do you draw from each of these tests? Give reasons for your answers.

(5)

(v) What is the relationship between the paired *t* value of part (i) and the *F* value for "week" of part (ii)?

(2)

2. The following data are from a two-factor experiment on sugar beet. The two factors are nitrogen (0 kg, 50 kg and 100 kg sulphate of ammonia per acre) and depth of winter ploughing (8 cm and 12 cm).

Yield of sugar (kg per acre) per treatment combination

Nitrogen	*0 kg*	*50 kg*	*100 kg*	*0 kg*	*50 kg*	*100 kg*
Depth of ploughing	*8 cm*	*8 cm*	*8 cm*	*12 cm*	*12 cm*	*12 cm*
	1054	1218	1406	1177	1374	1554
	1099	1258	1423	1160	1350	1572
	1080	1279	1468	1151	1351	1536
	1093	1273	1430	1145	1362	1561
Means	1081.50	1257.00	1431.75	1158.25	1359.25	1555.75

The analysis of variance (ANOVA) table is

Source	DF	SS	MS
Nitrogen	2	****	****
Depth	****	****	****
Nitrogen x Depth	****	2237.3	****
Error	****	****	397.9
Total	23	629744.5	

(i) Complete the ANOVA table and use it to assess what evidence the experiment provides regarding the six treatment combinations.

(12)

(ii) Draw a simple diagram using the six mean values which illustrates the effects of nitrogen, ploughing depth and their interaction.

(4)

(iii) Summarise your conclusions in non-technical language that the experimenter would understand.

(4)

3. A group of astronomers carried out a study of the relationship between light intensity and surface temperature. Data gathered on 24 stars in the cluster CYG OB1 are given in the table below. Note that there are three outlying points indicated by an asterisk (*).

Log surface	*Log light*	*Log surface*	*Log light*	*Log surface*	*Log light*
temperature	*intensity*	*temperature*	*intensity*	*temperature*	*intensity*
(x)	*(y)*	*(x)*	*(y)*	*(x)*	*(y)*
4.37	5.23	4.56	5.74	4.23	3.94
4.26	4.93	4.56	5.74	4.23	4.18
4.30	5.19	4.46	5.46	4.29	4.38
3.48*	6.05	4.57	5.27	4.42	4.42
4.26	5.57	4.37	5.12	4.42	4.18
3.49*	5.73	4.43	5.45	3.49*	5.89
4.48	5.42	4.43	5.57	4.29	4.22
4.29	4.26	4.42	4.58	4.49	4.85

* indicates outlying point

(i) A regression analysis of the full data set was performed using a statistical package and produced the following output.

```
The regression equation is
Log (light intensity)  =  7.74  –  0.628 × Log (surface temperature)

Predictor    Coeff     St dev      t        p
Constant      7.74      1.73      4.48   <0.001
Slope       –0.628     0.403     –1.56    0.134     s = 0.6207     R^2 = 9.9%
```

A quick look at the data suggests that there is a positive relationship between surface temperature and light intensity. However, the estimate of the slope is negative. Why is this?

(4)

(ii) The astronomers decided to asses the impact of the three outlying data points by deleting them and then calculating the following summaries.

$$\sum x = 92.13 \qquad \sum y = 103.70 \qquad \sum x^2 = 404.4303$$

$$\sum y^2 = 519.0608 \qquad \sum xy = 455.7101$$

Estimate the slope of the regression line after removing the outlying points and test the hypothesis that the slope is zero.

(8)

(iii) Construct a scatter plot of the full data set. Explain why the estimated slopes in (i) and (ii) have different signs.

(4)

(iv) What conclusions do you draw from these analyses?

(4)

4. Describe the circumstances in which simple exponential smoothing may be used to provide forecasts of future values of a time series X_t.

(4)

(i) If $\hat{x}(t,1)$ denotes the one-step-ahead forecast at time t where

$$\hat{x}(t,1) = \alpha x_t + \alpha(1-\alpha)x_{t-1} + \alpha(1-\alpha)^2 x_{t-2} + \alpha(1-\alpha)^3 x_{t-3} + \ldots$$

show that $\hat{x}(t,1) = \alpha\{x_t - \hat{x}(t-1,1)\} + \hat{x}(t-1,1)$.

(4)

(ii) The following data give the weekly sales of a commodity over a 12 week period where no special sales promotions or higher than usual levels of advertising took place.

Week	1	2	3	4	5	6	7	8	9	10	11	12
Sales	5105	5618	5514	5423	6044	5790	5437	5848	6253	5835	5740	6063

(a) Plot the data as a time series.

(2)

(b) Use simple exponential smoothing with $\alpha = 0.25$ and the first week's sales as the forecast for the second week's sales to forecast one week ahead at each time point.

(4)

(c) Plot the set of forecasts on the same graph as the original series and comment on their adequacy.

(3)

(d) Suppose a similar set of forecasts had been calculated for some other value of α. How would you compare the adequacy of the two sets of forecasts?

(3)

5. A biologist is studying the proportion of foetuses found dead in rat litters that each contain exactly five rats. Data are available from 169 litters and a summary of the observations is as follows.

Number of foetal deaths per litter	0	1	2	3	4	5
Number of litters	75	44	36	7	4	3

(i) Using n and p to denote the number of Bernoulli trials involved and the probability of a "success" respectively, write down expressions for the mean and variance of the number of successes assuming that it follows a binomial distribution.

(2)

(ii) Taking the death of an individual foetus as a "success", obtain an estimate of p from the data, and hence calculate the mean and variance of a binomial distribution for the number of successes.

(3)

(iii) Again taking the binomial distribution as a model for these data, calculate the expected numbers of litters having 0, 1, 2, ≥ 3 foetuses found dead. Draw a suitable diagram displaying the observed and expected numbers of litters having 0, 1, 2, ≥ 3 foetuses found dead, and test the goodness of the fit of the expected frequencies to the observed frequencies.

(8)

(iv) What assumptions about the individual foetuses within a litter, and the probabilities of death across litters, are implied by the model used in (ii)? Comment on how reasonable these assumptions are for these data. The comments should involve the sample variance $s^2 = 1.3273$ (**N.B.** this is not the variance of the binomial distribution).

(5)

(v) How would the analysis be changed if "success" was taken to be an individual foetus living rather than dying? [There is no need to carry out further calculations, but the required changes should be stated clearly, giving reasons where appropriate.]

(2)

6. The survival time of an individual after diagnosis of a certain fatal illness is T which is assumed to be exponentially distributed with probability density function

$$f\left(t\right)=\lambda e^{-\lambda t} \qquad t\geq 0 \ .$$

(i) Show that the probability that an individual survives after diagnosis for at least a time t_0 is given by the survivor function

$$S\left(t_0\right)=P\left(T\geq t_0\right)=e^{-\lambda t_0} \ .$$

(6)

(ii) The survival times (in days) of a group of 12 patients recruited into a study of this illness are given in the following table.

1327	1464	241	1027	20	332
308	20	100	71	889	229

Write down the likelihood function and obtain the maximum likelihood estimate of λ and an approximate value for its variance.

(10)

(iii) Rescale λ so that it is based on years rather than days. Using this rescaled value for λ, estimate the probability of surviving for at least one year.

(4)

7. (a) (i) Explain what is meant by the *non-response problem* in a sample survey, giving examples of where it might arise.

(5)

(ii) What steps might be taken to reduce the level of non-response?

(5)

(b) A survey organisation was interested in seeing what proportions of managing directors of companies in different sectors of business expected prospects for their company to improve in the next six months. The results of a survey of two sectors are given in the following table.

		Improve		
		No	*Yes*	Total
Sector	*Light engineering*	58	67	125
	Banking and financial services	74	126	200

(i) Calculate an approximate 99% confidence interval for the difference in proportions.

(7)

(ii) Do the results suggest that the two sectors have the same views? Give reasons for your answer.

(3)

8. The following data (x) are the number of plants per square metre obtained in an experiment where a particular type of grass seed was applied at four rates. The experiment was laid out as a completely randomised design.

	Rate of application			
	1	*2*	*3*	*4*
	29	180	332	910
	13	90	444	880
	21	120	190	460
Mean	21	130	322	750
Variance	64	2100	16204	63300

(i) State the assumptions required for a one-way analysis of variance (ANOVA) and whether these data appear to satisfy these assumptions. Give reasons.

(4)

(ii) The following three transformations have been suggested for these data: $\sqrt{x}$, $\log_e(x)$ and $(1/x)$. The following table contains values for the means and variances of the transformed data although some are missing. Complete the table and then say which of the transformations should be performed prior to carrying out a one-way ANOVA. Give reasons.

(6)

Rate	**Transformation**					
	$\sqrt{x}$		$\log_e(x)$		$1/x$	
	mean	*variance*	*mean*	*variance*	*mean*	*variance*
1	4.52	0.79	2.992	0.1630	0.05301	****
2	11.29	3.94	****	****	0.00833	0.0000077
3	17.69	****	5.716	0.1861	0.00351	0.0000025
4	27.09	23.96	6.575	0.1479	****	0.0000004

(iii) Perform the ANOVA for the transformation chosen. Sums of squares for the various transformations are:

SS	x	$\sqrt{x}$	$\log_e(x)$	$1/x$
Rates	928778	830.78	21.153	0.0053826
Total	1092114	915.16	22.389	0.0063478

What conclusions do you draw from the ANOVA?

(6)

(iv) Calculate a 95% confidence interval for the rate 3 mean value on the transformed scale. Back-transform this interval onto the original scale.

(4)