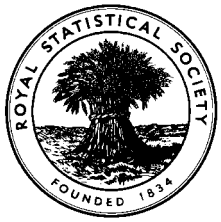
EXAMINATIONS OF THE ROYAL STATISTICAL SOCIETY
(formerly the Examinations of the Institute of Statisticians)



HIGHER CERTIFICATE IN STATISTICS, 1998

Paper III : Statistical Applications and Practice

Time Allowed: Three Hours

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

All questions carry equal marks.

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 $\binom{n}{r}$ *is the same as* ${}^{n}C_{r}$ *and that* $\ln$ *stands for* $\log_e$.

1. Six cartons were sampled randomly from each of four batches of milk and were stored for 14 days. After this time they were tested for the presence of bacteria which cause spoilage. The bacterial counts for samples are given in the table below together with the mean and variance for each batch.

Batch	*Bacterial count*	*Mean*	*Variance*
A	439, 480, 497, 464, 460, 462	467.000	387.200
B	207, 204, 186, 186, 208, 204	199.167	106.567
C	74, 71, 72, 87, 89, 75	78.000	62.400
D	252, 227, 234, 258, 260, 259	248.333	203.467

(i) Make a plot of variance against mean for the four batches. What does this plot reveal about the need for a transformation if the intention is to perform an analysis of variance of the results of the investigation? In particular, which assumption required for the validity of the analysis appears to be violated?

The table below gives the square roots of the counts, together with their sample means and sample variances.

Batch	*Square root of count*	*Mean*	*Variance*
A	20.95, 21.91, 22.29, 21.54, 21.45, 21.49	21.6050	0.206550
B	14.39, 14.28, 13.64, 13.64, 14.42, 14.28	14.1083	0.134817
C	8.60, 8.43, 8.49, 9.33, 9.43, 8.66	8.8233	0.193427
D	15.87, 15.07, 15.30, 16.06, 16.12, 16.09	15.7517	0.205577

(ii) Explain why the information contained in this table indicates that the transformation has made the data conform more closely to the assumption which was violated by the un-transformed data.

(iii) Perform an analysis of variance of the transformed data to assess whether there is evidence of differences between the bacterial content of the batches. Note that since the square root transformation has been applied the sum of squares of the transformed data is equal to the sum of the un-transformed counts.

Turn over

2. An investigation was carried out into the levels of chlordanes in wild polar bears. The levels of chlordanes in 25 sub-adult polar bears, 23 adult females and 20 adult males are given in the table below, together with the corresponding transformed values obtained by taking logarithms to the base 10 to two places of decimals. The bears were those which the investigators were able to locate and anaesthetise for long enough to collect tissue samples which could be assayed for chlordanes and other organo-chlorine compounds. It is safe to assume that the bears located were a random sample of those in the areas covered by the investigation.

Chlordane levels in tissue (nano-grams/gram)

Sub-adults		**Adult females**		**Adult males**	
Original Data	*Log transformed*	*Original data*	*Log transformed*	*Original Data*	*Log transformed*
1081	3.03	813	2.91	475	2.68
1188	3.07	881	2.95	610	2.79
1449	3.16	1170	3.07	675	2.83
1657	3.22	1219	3.09	788	2.90
1805	3.26	1350	3.13	916	2.96
1846	3.27	1409	3.15	1089	3.04
1960	3.29	1521	3.18	1321	3.12
2069	3.32	1932	3.29	1330	3.12
2089	3.32	2049	3.31	1357	3.13
2272	3.36	2201	3.34	1584	3.20
2664	3.43	2771	3.44	1802	3.26
2971	3.47	2859	3.46	1877	3.27
3329	3.52	2985	3.47	2045	3.31
3900	3.59	3094	3.49	2282	3.36
4349	3.64	3855	3.59	2520	3.40
4539	3.66	3876	3.59	2888	3.46
4746	3.68	4142	3.62	3060	3.49
4819	3.68	4397	3.64	4994	3.70
5597	3.75	4832	3.68	6238	3.80
5607	3.75	5473	3.74	6241	3.80
6013	3.78	5825	3.77		
6291	3.80	6414	3.81		
7908	3.90	7511	3.88		
8219	3.91				
8505	3.93				

Note that the observed values have been ordered.

(Question continued on next page)

(i) Make box-plots of the raw data and also of the logarithms to the base 10 of the observed chlordane levels.

(ii) Describe how the shapes of the distributions of the raw data and the log transformed data differ.

(iii) There is evidence from studies of other animals that chlordanes tend to accumulate most in fatty tissue and can be passed from female mammals to their offspring through the mother's milk during the nursing period. Females tend to have considerably more fatty tissue than males. The investigators are interested in comparing average levels of chlordanes between sub-adults and male and female adults. A partially completed analysis of variance table for the log transformed data is given below. Complete this analysis and explain in non-technical terms what can be concluded from the analysis.

Analysis of Variance of Chlordane levels in polar bears.

Source of Variation	*Sum of squares*	*Degrees of freedom*	*Mean square*	*Variance ratio*
Between groups	0.8930			
Residual				
Total	6.3774	67		

(iv) Carry out a follow up analysis by calculating individual 95 per cent confidence intervals for the three groups, using the pooled estimate of error variance in determining the standard errors of the group means. Explain what this analysis reveals.

(v) Back transform the group means and the 95 percent confidence intervals to the original scale.

Turn over

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

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

(b) A survey organisation defines the "true level of business confidence" for a particular sector of economic activity as the proportion of managing directors of all companies in that sector who expect prospects for their company to improve in the next six months.

In the Light Engineering sector the managing directors of 67 out of a random sample of 125 companies stated that they expected prospects for their company to improve in the next six months. For a random sample of 200 companies in the Banking and Financial Services sector the corresponding figure was 126 managing directors reporting that they expected prospects for their company to improve over the same period.

(i) Do these results provide evidence of a difference between true levels of business confidence (proportions expecting an improvement) in the two sectors?

(ii) Calculate a 95 percent confidence interval for the difference between the true levels of business confidence in the two sectors.

(iii) A business analyst wants to calculate a 95% confidence interval for the level of business confidence in the Light Engineering sector. If the true level is 0.6, what sample size would be required to produce a 95% confidence interval which has a width of 0.08?

4. Water uptake of seeds of a particular variety of wild oats was measured at 2 hourly intervals for a period of 24 hours. The uptake is measured in grams per gram of dry weight of seed. For a particular series of experiments the results are shown in the table below.

Time (t)	*Water uptake (y)*	*w (= t/y)*
2	0.69	2.883
4	1.50	2.664
6	1.43	4.180
8	1.64	4.866
10	2.01	4.952
12	1.96	6.122
14	2.28	6.133
16	2.10	7.620
18	2.02	8.887
20	2.33	8.578
22	2.32	9.454
24	2.42	9.936

Theory suggests that water uptake is expected to be related to time by the hyperbolic law:

$$y = \frac{ct}{d+t}.$$

(i) Show that if the hyperbolic law does hold then there is a linear relationship between the transformed variable w and t, where $w = t/y$.

(ii) Plot the values of w (which are given in the third column of the table), against t.

(iii) Use the method of least squares to determine estimates $\hat{\alpha}$ and $\hat{\beta}$ of the parameters α and β of the best fitting straight line:

$$w = \alpha + \beta t\,.$$

[Note that $\sum w = 76.39073$ and $\sum wt = 1190.46022.$]

(iv) Hence, obtain estimates $\hat{c}$ and $\hat{d}$ of the parameters c and d in the hyperbolic law.

(v) What value of water uptake does the fitted hyperbolic law predict at 16 hours?

(vi) Indicate briefly any technical difficulties which arise in obtaining standard errors of $\hat{c}$ and $\hat{d}$ when you are given the standard errors of $\hat{\alpha}$ and $\hat{\beta}$.

Turn over

5. (a) The January prices and volumes, that is the actual number of shares traded in thousands of shares, for four companies are shown for 1995 and 1997 in the table below.

	A		**B**		**C**		**D**	
	Price	*Volume*	*Price*	*Volume*	*Price*	*Volume*	*Price*	*Volume*
Jan 1995	£2.98	229.7	£4.40	484.3	£3.85	137.8	£11.41	2721.5
Jan 1997	£2.45	167.3	£6.43	777.3	£2.66	165.2	£15.15	3193.7

(i) Calculate a current weighted Price index (Paasche index) for January 1997 using the volumes as weights.

(ii) Express the January 1997 share price of company D as a percentage of its January 1995 price.

(iii) Explain why it is not merely a coincidence that the values obtained in (i) and (ii) are so close.

(b) The value of a particular share (in £ sterling) at close of trading on each of 10 consecutive trading days is shown in the table below.

Day	*Share value*	*Day*	*Share value*
1	3.85	6	3.43
2	3.56	7	3.28
3	3.65	8	3.43
4	3.54	9	3.35
5	3.71	10	3.45

(i) Plot these data.

(ii) Use the method of exponential smoothing with a smoothing parameter of 0.4 to provide one step ahead predictions for each day after day 1 and plot the predictions on the same graph.

(iii) Explain briefly, without doing any further calculations, what would be the effect on the predictions of using a smoothing parameter of 0.9.

6. The values of the daily change in the FTSE-100 share index were calculated for the period from 31 December 1994 to 29 December 1996. Summary statistics for these daily changes are given below

n	*mean*	*median*	*Standard Deviation*
520	0.52	0.35	23.43

The table below gives a grouped frequency distribution for the daily changes and some expected frequencies which have been calculated by assuming that the daily changes have a normal distribution.

(i) Complete the calculation of the expected frequencies and then test whether the assumption of normality is a reasonable one.

Grouped frequency distribution of daily changes in the FTSE-100 share index

Daily change	*Observed Frequency*	*Expected Frequency*
–67.5 or less	3	
–67.5 to –52.5	4	
–52.5 to –37.5	26	21.063
–37.5 to –22.5	54	57.512
–22.5 to –7.5	90	105.632
–7.5 to +7.5	147	
+7.5 to +22.5	102	108.573
+22.5 to +37.5	66	
+37.5 to +52.5	25	22.873
+52.5 to +67.5	2	5.789
+67.5 or more	1	1.106
Total	520	

(ii) In many investigations the number of observations is much smaller, for example in many designed experiments the number of observations may be less than 30. Why would it be uninformative to use the method of part (i) to assess the normality of the residuals in such a designed experiment?

(iii) Describe a graphical method which can be used to assess the normality of a set of residuals in a designed experiment. Illustrate your answer by showing how the method could be used for a randomised block design, with five treatments compared in six blocks?

Turn over

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

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

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

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

(ii) The survival times in days, of a group of 12 patients recruited into a study of this illness are given in the table below. Obtain the maximum likelihood estimate of λ.

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

(iii) Obtain an approximate value for the variance of the maximum likelihood estimate of λ.

(iv) Suppose it was later found that, in addition to the 12 patients described above (who all died), there were another 3 who did not die during the follow up period of the study. At the end of the follow up period these three patients had survived for 641 days, 234 days and 87 days, respectively. Combine this information with the data given in part (ii) to obtain the maximum likelihood estimate of λ.

8. (i) Explain what is meant by a factorial experimental design and state what advantages such a design has over a number of separate single factor experiments.

The Journal of Applied Ecology for 1997 reports the results of a study of the feeding behaviour of goats, red deer and South American camelids on three different mixtures of Scottish plants. A group of fifteen of each type of animal was randomly divided into groups of five. One group of each type of animal was then allocated to feed on a plot of land where one of the three plant mixtures was growing. The three plant mixtures were:

sown grasses,
natural grasses,
heathers.

The mean organic matter intake of each group of animals (kg per day) is given in the table below.

Plant mixture	*Goats*	*Red Deer*	*Camelids*
Sown grasses	0.907	2.033	1.137
Natural grasses	1.348	2.437	1.973
Heathers	0.646	1.888	0.970

The total of the intake measurements for all 45 animals was 66.689 and the sum of the squares of the intake values for all 45 animals was 114.85.

(ii) Carry out an analysis of variance table to analyse the effects of animal types, plant mixtures and any interaction between these factors.

(iii) Construct a suitable graphical display to illustrate the presence or absence of interaction.