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

HIGHER CERTIFICATE IN STATISTICS, 1997

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

1. The quarterly profits, in thousands of dollars, of a small but growing company are shown in the table below. The table also includes the centred 4 point moving average of the time series and the differences (profit – centred moving average).

		Profit ($'000's)	*Moving average*	*Difference*
1992	Q1	45.5		
	Q2	59.3		
	Q3	82.8	67.900	
	Q4	69.4	75.163	-5.7625
1993	Q1	74.7	82.125	
	Q2	88.2	87.762	0.4375
	Q3	109.6	92.887	16.7125
	Q4	87.7	98.325	-10.6250
1994	Q1	97.4	102.637	-5.2375
	Q2	109.0		
	Q3	123.3	115.250	8.0500
	Q4	118.9	121.938	-3.0375
1995	Q1	122.2	129.587	-7.3875
	Q2	137.7	136.400	1.3000
	Q3	155.8		
	Q4	140.9	147.700	-6.8000
1996	Q1	144.3	152.475	-8.1750
	Q2	161.9	156.913	4.9875
	Q3	169.8		
	Q4	162.4		

(i) Complete the calculation of the moving averages and the differences.

(ii) Plot the data together with the moving average.

(iii) Estimate the seasonal effects assuming that an additive model is appropriate for describing the data.

(iv) Explain how the deseasonalized values are calculated, but do not calculate them.

(v) A straight line was fitted to the deseasonalized data and the result was

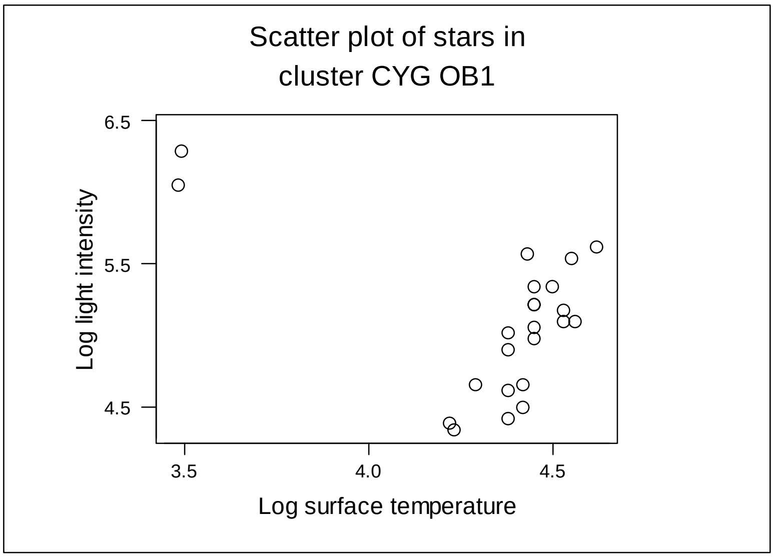
$$\text{profit} = 50 + 6\,t$$

where t is the number of quarters from the start of the series, so that, for example, the value of t for 1993 Q2 is 6.
Use this information concerning the fitted line and the seasonal effects to produce profit projections for 1997. What assumptions does this require?

Turn over

2. A group of astronomers carried out a study of the relationship between light intensity and surface temperature. Data gathered on 23 stars in the cluster CYG OB1 are given in the table below and are plotted in the scatter plot.

Log surface temperature (x)	*Log light intensity (y)*	*Log surface temperature (x)*	*Log light intensity (y)*
4.38	5.02	4.53	5.10
4.42	4.66	4.45	5.22
4.29	4.66	4.53	5.18
4.38	4.90	4.43	5.57
4.22	4.39	4.38	4.62
3.48	6.05	4.45	5.06
4.38	4.42	4.50	5.34
4.56	5.10	4.45	5.34
4.45	5.22	4.55	5.54
3.49	6.29	4.45	4.98
4.23	4.34	4.42	4.50
4.62	5.62		



(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 = 8.41 - 0.763 Log surface temperature

Predictor      Coef      Stdev      t-ratio      p

Constant      8.410      1.512        5.56     0.000
slope        -0.7628     0.3469      -2.20     0.039                s = 0.4712     R^2 = 18.7%
```

Explain why the slope of the line is negative.

(Question continued on next page)

(ii) After removing two of the data points a further regression analysis produced the following output:

The regression equation is   Log light intensity = - 8.48 + *** Log surface temperature

```
Predictor     Coef      Stdev    t-ratio      p

Constant     -8.475      2.484     -3.41    0.003
slope         ***       0.5604     ***      0.000        s = 0.2557      R^2  = 60.7%
```

The value of the slope, which is positive, has been deleted from the output.

Which points were removed? Calculate the value of the slope which is missing from this output.

Note: for the full set of 23 stars

$$\sum x = 100.04, \quad \sum y = 117.12, \quad \sum x^2 = 436.9760,$$
$$\sum y^2 = 602.1320, \quad \sum xy = 508.0134.$$

Explain why the removal of two points from the data set causes such a marked change in the results of the analysis.

(iii) How would you advise someone who wanted to use either of the equations as a summary of the relation between light intensity and surface temperature for the stars in the cluster CYG OB1?

3. Write an essay describing the use of residuals as diagnostic tools in data analysis. You should define what residuals are and illustrate your explanation with sketches of the types of patterns one might expect to see when examining residual plots and indicate what the patterns reveal.

Turn over

4. Silicon chip manufacture involves a complex process in which layers are deposited on a wafer of silicon by vapour deposition at high temperatures. An electronics engineering team investigated the effect of deposition time and deposition temperature on the thickness of a particular layer. They particularly wanted to find out whether either factor alone could be used to control the thickness of the layer. They used two settings of deposition time (Low and High relative to currently used time) and two temperatures (1210°C and 1240°C). Measurements of the thickness of the layer were made on each of five silicon wafers produced under each set of conditions. The measurements are given in the table below.

Thickness of layer (µm)

		Deposition Time	
		High	*Low*
Deposition Temperature	1210	14.90, 14.69, 14.52, 15.14, 14.63.	13.78, 14.18, 13.58, 13.58, 13.81.
	1240	14.49, 14.33, 13.94, 14.31, 14.18.	14.27, 14.37, 14.16, 14.03, 14.20.

The sum of the 20 measurements is 285.09 and the corresponding sum of squares is 4067.00.

(i) Perform an analysis of variance, including tests for the effects of time and temperature and the interaction between time and temperature.

(ii) Produce a diagram of means and their standard errors which makes clear the nature of any interaction that there may be between time and temperature.

(iii) Write a short report (4 or 5 informative sentences) which explains the findings of your analysis in non-technical language for the team who carried out the investigation.

5. You have been asked by a horticulturalist at a research station to advise on the design of an experiment on the growth of tomatoes which is to be undertaken in a glasshouse environment. The experiment is to compare the effects on mean fruit yield of four different nutrient solutions.
Plants will be grown in commercial “grow-bags”, which are rectangular plastic bags containing a standard growing medium. There will be four plants to a bag and there will be the same number of plants per nutrient solution. It is not possible to provide different nutrient solutions to plants which are in the same bag. There is enough space in the glasshouse for two rows each containing 32 full grow bags, laid side by side.

   (i) The horticulturalist has heard of completely randomised and randomised blocks designs and asks for your advice on which one to use. What questions would you ask the horticulturalist before making a recommendation to him?

   (ii) Explain what constitutes an experimental unit in this experiment and what outcome variable would form the basis of your analysis.

   (iii) Indicate, for both designs, how the nutrient solutions would be allocated to plants.

   (iv) Give a breakdown of the sources of variation and the degrees of freedom for the analysis of both designs.

6. A health questionnaire was administered to people in five areas of a country, the selection being made randomly within areas. One question asked the respondent to rate their own health as "good", "fair" or "poor". The results are summarised in the table below.

Area	*Good*	*Fair*	*Poor*	*Total*
Ruthven	459	178	43	680
Mossmont	926	506	103	1535
Windgyle	954	442	78	1474
Dundonan	985	507	85	1577
Ainster	365	176	35	576
Total	3689	1809	344	5842

   (i) Carry out an appropriate statistical test of whether there are differences between perception of health in different areas and investigate the nature of these differences.

   (ii) Do you feel that the survey will provide an accurate comparison of health status between the areas? Suggest an alternative way of measuring health status, stating any advantages and disadvantages compared to the survey described here.

Turn over

7. The table below gives the failure times in hours of two makes of equipment. The data were gathered in an attempt to answer the question of whether the two types differ in average failure time.

   (i) Draw box and whisker plots of the two sets of observations and describe briefly what they indicate about the shape of the underlying distributions from which the data have been sampled.

   (ii) When testing whether two populations differ in location one may use a parametric test such as the *t*-test or a non-parametric test such as the Mann-Whitney test. Carry out an appropriate test to answer the question of whether the two types of equipment differ in average failure time and explain your choice of test. Describe precisely the null hypothesis that is being tested by the test you choose.

Type A	*Type B*
171, 257, 288, 295, 396, 397, 431, 435, 554, 568, 795, 902, 958, 1004, 1104, 1212, 1283, 1378, 1621, 2415	212, 236, 262, 272, 286, 311, 336, 340, 412, 446, 449, 670, 686, 786, 811, 836, 936, 978, 1335, 1678
sum = 16464 sum of squares = 19648218	sum = 12278 sum of squares = 10544940

8. In a pilot study of anaesthesia in cardiac surgery twenty two patients were randomly assigned to three treatment groups. The table shows the red cell folate levels in the three groups after 24 hours of treatment.

	Group A (*n* = 8)	*Group B* (*n* = 9)	*Group C* (*n* = 5)
	243	206	241
	251	210	258
	275	226	270
	291	249	923
	347	255	328
	354	273	
	380	285	
	392	295	
		309	
Sum	2533	2308	2020
Sum of squares	826145	602898	1157058

(i) Make an informative plot of the data. One of the data points seems to be unusual compared with the rest. State which one it is and why you regard it as unusual.

(ii) A one-way analysis of all the data performed by a computer package is summarised in the analysis of variance table below. Some values have been left out of the table. Complete the table and state the conclusions that you feel are justified by this analysis.

Source of variation	*Sum of squares*	*Degrees of freedom*	*Mean square*	*Variance ratio*
Between treatments	*	2	*	*
Residual	*	19	19797	
Total	446405	21		

(iii) After some discussion with the surgical team who conducted the study you decide to leave out the observation which is unusual because it is not possible to determine whether it is a genuine observation or a mistake in recording the data.
Re-analyse the data and draw up an analysis of variance table similar to the one above. State the conclusions which are justified by your second analysis.

(iv) How would you explain any difference between your conclusions from the two analyses?