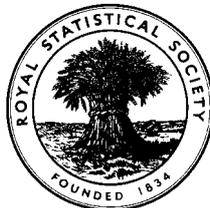**EXAMINATIONS OF THE ROYAL STATISTICAL SOCIETY**

(formerly the Examinations of the Institute of Statisticians)

HIGHER CERTIFICATE IN STATISTICS, 2000

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

There is a worksheet for use with question 8.

If you answer question 8, you must hand the worksheet in with your answer book. ENSURE YOU HAVE WRITTEN YOUR CANDIDATE'S NUMBER ON THE WORKSHEET.

1

1. As part of an investigation of the immunology of a certain type of inflammatory skin reaction, skin biopsies were taken from a group of 150 randomly chosen subjects whose reaction to the tuberculin antigen was of the Listeria (L) type and from another group of 200 subjects who showed the Koch (K) type reaction to the antigen. The result of a particular immunological reaction on each skin biopsy was classified as positive (+) or negative (−).

   (i) Given the results below, test whether the proportion of + test results differs significantly between the two groups of subjects.

		Reaction type	
		K	L
Immunological	+	127	64
test result	−	73	86

   (7)

   (ii) Determine a 95% confidence interval for the difference between the proportions of positive test results in subjects.

   (6)

   (iii) What conclusions do you draw from your confidence interval?

   (3)

   (iv) Why does a confidence interval give a "better" answer than a point estimate?

   (2)

   (v) Assuming that the proportions of positive tests in the two groups were unchanged, approximately how large a sample size would be needed to produce a confidence interval which had a width that was one quarter of the width of the interval obtained in (ii)?

   (2)

2. A government agency has developed a formula for calculating the theoretical carrying capacity of an urban roundabout. Owing to a controversy about the usefulness of the formula, the civil engineering department of a university was asked to investigate the situation. This they did by selecting ten roundabouts to which the formula was applicable, calculating the theoretical carrying capacity by the formula and observing the actual capacity for each roundabout. The results of this investigation are shown below.

Carrying capacity (hundreds of vehicles per hour)			
Theoretical capacity by formula (x)	*Observed capacity (y)*	*Theoretical capacity by formula (x)*	*Observed capacity (y)*
32.0	33.0	40.0	45.0
33.0	38.4	43.0	43.2
34.0	37.4	45.0	49.0
35.1	42.0	46.9	50.0
36.9	39.4	48.0	47.4

$$\sum x^2 = 15840.23, \qquad \sum y^2 = 18317.68, \qquad \sum xy = 17005.66.$$

The method of least squares was used to fit a straight line to these data and some of the output from the package is shown below.

```
The regression equation is
y =  ***** + ***** x

Predictor      Coef      StDev
Constant       *****     5.114
x              *****     0.1285

S = 2.315      R-Sq = 84.3%

Analysis of Variance

Source            DF        SS         MS        F      P
Regression        1       229.31     229.31   42.80   0.000
Error             8        42.87       5.36
Total             9       272.18
```

(i) Plot the data. (4)

(ii) Calculate the missing values ***** in the output above (estimates of the parameters of the line) and draw the fitted line on your scatter plot. (8)

(iii) What information is provided by the value labelled as "R-Sq" in this output? (3)

(iv) Calculate 95% confidence intervals for the slope and the intercept. (5)

3. In an investigation concerned with storage techniques associated with bone marrow transplantation, samples of bone marrow, all from the same subject, were stored at $-70°C$ for various lengths of time. After return to $37°C$, the viability of each sample was determined. Viability is a measure of the ability of the bone marrow cells to function naturally. The results of the viability tests are given below; a high value indicates high viability.

	Storage time					
	0 hours	*24 hours*	*36 hours*	*48 hours*	*72 hours*	*120 hours*
	150	151	150	145	132	74
	140	135	64	132	114	102
	125	123	136	119	142	91
Total	415	409	350	396	388	267

Sum of all viability results = 2225.
Sum of squares of all viability results = 286327.

(i) Perform an appropriate one-way ANOVA analysis to assess whether storage time affects viability.

(12)

(ii) If the sample at 36 hours with viability 64 is omitted from the analysis, the residual sum of squares becomes 1948. Construct a revised analysis of variance table and a table of means appropriate to this situation. What is the conclusion indicated by the revised analysis? How would you explain this to the investigator who provided the data?

(8)

5

4. A botanist investigating the dynamics of the formation of colonies of various species of clover performed the following experiment. Thirty-two plots were sown as follows: 8 plots for each of species A, B, C, D. Since the intention of the experiment was to examine how the various species developed as time progressed, two plots for each species were randomly allocated to each of the times (i) 9 weeks, (ii) 12 weeks, (iii) 15 weeks, (iv) 18 weeks after sowing. At these times the chosen plots were examined and the leaf area per unit ground area was determined for each plot. The results are given below.

		Weeks from sowing			
		9	*12*	*15*	*18*
Species	*A*	0.8, 1.3	3.6, 3.1	3.7, 3.3	1.9, 1.2
	B	0.3, 0.8	2.5, 2.9	4.7, 3.9	6.5, 5.2
	C	0.5, 0.7	2.4, 2.5	2.8, 3.1	1.5, 1.3
	D	0.2, 0.4	1.0, 1.3	1.6, 1.9	1.8, 1.9

These data were analysed using a computer package and part of the output is shown below.

Analysis of Variance for leaf area per unit ground area

Source	DF	SS	MS
Species	*	18.8012	6.2671
Time	3	*****	******
Species×Time	*	21.0862	******
Error	*	******	******
Total	31	70.7987	

(i) Complete the analysis of variance table and use it to assess whether the experiment provides evidence that the four species behave differently with time.

(10)

(ii) Draw a suitable diagram to illustrate any species-time interaction that there may be.

(5)

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

(5)

5.   The Poisson distribution may be used to provide a simple statistical model for the number of vehicles passing a traffic survey point when traffic is light. The data below were collected from 120 independent one-minute intervals at a survey point.

Number of vehicles	0	1	2	3	4	5	6	7	8	≥ 9
Number of intervals	8	30	32	20	13	9	5	2	1	0

(i)   Assuming that a Poisson model is appropriate, calculate the expected numbers of intervals with x vehicles per minute for $x = 0, 1, 2, \ldots, 8$ and for $x \geq 9$.

(7)

(ii)   Use an appropriate statistical test to assess whether a Poisson model is reasonable for this set of data.

(7)

(iii)   Assuming that a Poisson model is valid, obtain approximate 95% confidence intervals for:

(a)   the mean number of vehicles passing the survey point per minute,

(b)   the probability of at least one vehicle passing the survey point in a minute.

(6)

Turn over

6. The central surface brightness was measured for 68 galactic globular clusters of stars and the results were related to the degree of concentration of the cluster which was categorised as low, medium or high. Descriptive statistics for these data are given below.

Variable: central surface brightness

Concentration of cluster	N	Mean	Median	St. Dev.	Min.	Max.	Lower quartile	Upper quartile
Low	20	21.7	20.75	2.23	18.9	25.7	20.025	23.975
Medium	26	18.1	17.85	2.30	14.4	24.0	16.600	19.250
High	22	17.3	16.30	3.14	14.1	25.2	15.275	17.650

(i) Draw box and whisker plots side by side on one sheet of graph paper for the three categories of globular cluster.

(9)

(ii) Explain carefully what the plots suggest about the distribution of central surface brightness and the way in which central surface brightness is related to concentration of the cluster.

(6)

(iii) Suppose you wanted to assess the evidence that central surface brightness is related to concentration of the cluster. Give three reasons why a one-way analysis of variance of the original values might not be appropriate.

(5)

7. A comparison of the reliability of engine bearings made from different alloys was performed by testing ten bearings of each type. The times to failure in units of millions of cycles are given below.

Alloy A	Alloy B
5.30	3.19
5.53	4.26
5.60	4.47
6.30	4.53
6.92	4.67
12.51	4.69
12.95	6.79
16.04	9.37
18.21	12.75
18.24	12.78

A computer package produced the five-number summaries in the form

    (minimum, lower quartile, median, upper quartile, maximum)

giving

    Alloy A   (5.30, 5.58, 9.71, 16.58, 18.24)

    Alloy B   (3.19, 4.42, 4.68, 10.22, 12.78)

(i)   Draw dot plots, side by side, of the results for the two alloys and mark the mean failure time for each alloy.

(6)

(ii) Use a non-parametric test to assess whether the distribution of times to failure differs for the two types of alloy. Explain carefully why your test might be more appropriate than a two sample t-test.

(10)

(iii) State clearly the conclusion that can be drawn from the analysis.

(4)

Turn over

8. The table below gives the total deaths from lung diseases in the UK for each quarter for the years 1974 to 1979.

1974	1	8291	1977	1	8059
	2	6223		2	6155
	3	4841		3	4766
	4	6785		4	6393
1975	1	8760	1978	1	8779
	2	6093		2	6046
	3	4548		3	4552
	4	6700		4	6492
1976	1	9857	1979	1	9386
	2	5227		2	5347
	3	4145		3	4259
	4	6489		4	6206

Plot the data as a time series and describe the main features of the data.

(7)

The data are also given on the worksheet for this question together with an incomplete set of centred four point moving average values and an incomplete set of differences between the actual deaths and the corresponding moving average values. Complete the calculation of moving average values and differences.

(5)

Complete the calculation of seasonal components on the assumption of an additive model (trend plus seasonal plus residual) for the data.

(4)

The trend line for the data is

$$\text{deaths} = 6831 - 31.8t$$

where t = number of quarters since the start of 1974 (for example, for quarter 2 of 1976, $t = 10$).

Use this information to predict deaths for 1980.

(4)

> **The worksheet for use with this question is supplied as a separate sheet of paper.**
>
> **If you answer this question, you must hand the worksheet in with your answer book.**
>
> **ENSURE YOU HAVE WRITTEN YOUR CANDIDATE'S NUMBER ON THE WORKSHEET.**

_____
The worksheet is now attached as page 11 of this examination paper

Worksheet for UK lung disease deaths for use with Question 8.

		Y	M.Av.	$Y-$ M.Av.
1974	*1*	8291		
	2	6223		
	3	4841	6593.63	-1752.63
	4	6785		
1975	*1*	8760	6583.13	2176.88
	2	6093	6535.88	-442.88
	3	4548	6662.38	-2114.38
	4	6700	6691.25	8.75
1976	*1*	9857	6532.63	3324.38
	2	5227		
	3	4145		
	4	6489	6096.00	393.00
1977	*1*	8059		
	2	6155	6355.25	-200.25
	3	4766	6433.25	-1667.25
	4	6393	6509.63	-116.63
1978	*1*	8779		
	2	6046		
	3	4552	6543.13	-1991.13
	4	6492		
1979	*1*	9386	6407.63	2978.38
	2	5347	6335.25	-988.25
	3	4259		
	4	6206		

Calculation of seasonal effects

	Q1	*Q2*	*Q3*	*Q4*	
	2176.88	-442.88	-1752.63		
	3324.38		-2114.38	8.75	
detrended data		-200.25		393.00	
			-1667.25	-116.63	
	2978.38	-988.25	-1991.13		
seasonal means					
seasonal effects					