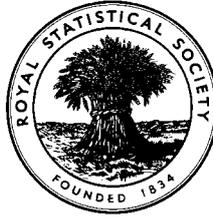**EXAMINATIONS OF THE ROYAL STATISTICAL SOCIETY**

(formerly the Examinations of the Institute of Statisticians)

GRADUATE DIPLOMA IN STATISTICS, 1999

Options Paper

Time Allowed: Three Hours

This paper contains four questions from each of six option syllabuses. Each option syllabus is one Section.

Section	A:	*Statistics for Economics*
	B:	*Econometrics*
	C:	*Operational Research*
	D:	*Medical Statistics*
	E:	*Biometry*
	F:	*Statistics for Industry and Quality Improvement*

Candidates should answer **FIVE** *questions chosen from* **TWO SECTIONS ONLY**.

Do **NOT** *answer more than* **THREE** *questions from any* **ONE** *Section.*

ANSWER EACH SECTION IN A SEPARATE ANSWER-BOOK.

Label each book clearly with its Section letter and name.

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.

This examination paper consists of nn printed pages. This front cover is page 1. The reverse of the front cover, which is intentionally left blank, is page 2. Question 1 of Section A starts on page 3. There are 24 questions altogether in the paper, 4 in each of the 6 Sections.

SECTION A - STATISTICS FOR ECONOMICS

A1. Statistics of investment in stocks and work in progress in year t (S_t) and of gross domestic product in year t (GDP_t), for $t = -10, -9, ..., 10$ corresponding to the years 1976 to 1996, in £m at 1990 prices relating to the United Kingdom, are compiled from Table 1.3 of *United Kingdom National Accounts*, 1997 edn.

$\Delta GDP_t \equiv GDP_t - GDP_{t-1}$. GDP for 1975 is used to obtain ΔGDP for 1976, so there are $n = 21$ observations on (S_t, ΔGDP_t).

It is found that $\sum S_t = 20,486$, $\sum (\Delta GDP_t) = 214,853$, $\sum (S_t)^2 = 185,372,000$, $\sum (\Delta GDP_t)^2 = 4,342,813,696$, and $\sum (S_t . \Delta GDP_t) = 707,880,192$.

(i) Estimate the regression model $S_t = \alpha + \beta(\Delta GDP_t) + u_t$ by ordinary least squares (OLS), together with the standard deviation of u_t and the standard errors of the estimates of α and β. What are the sums of the squares of S_t about (a) its mean and (b) the regression?

(7)

(ii) The model $S_t = \alpha + \beta_1 GDP_t + \beta_2 GDP_{t-1} + u_t$ is similarly estimated, as
$$S_t = -155 + 0.23324 \ GDP_t - 0.23585 \ GDP_{t-1}, \quad R^2 = 0.704, \quad s = 1651$$
$$\quad (2754) \ (0.03569) \quad\quad (0.03645) \quad\quad\quad\quad\quad rss = 49,037,296$$
with standard errors in parentheses.

Test the null hypothesis that $\beta_1 = -\beta_2$.

(3)

(iii) Adding the time variable to these models gives, using OLS,
$$S_t = -1422 + 0.23714 \ \Delta GDP_t - 58.25t, \quad R^2 = 0.716, \quad s = 1617$$
$$\quad (502) \ (0.03525) \quad\quad (58.83) \quad\quad rss = 47,052,340$$
and
$$S_t = -26949 + 0.26405 \ GDP_t - 0.21095 \ GDP_{t-1} - 605.8t, \quad R^2 = 0.790, \quad s = 1429$$
$$\quad (10391) \ (0.03302) \quad\quad (0.03292) \quad\quad\quad (228.7) \quad\quad rss = 34,707,524$$

while the same OLS process gives
$$S_t = -26949 + 0.05310 \ GDP_t + 0.21095 \ \Delta GDP_t - 605.8t, \quad R^2 = 0.790, \quad s = 1429$$
$$\quad (10391) \ (0.02159) \quad\quad (0.03292) \quad\quad\quad (228.7) \quad\quad rss = 34,707,524$$

Test whether the coefficients of t in these three estimated models are significantly different from zero. Discuss the inter-relationship between the five models. Which do you prefer? Find $\overline{R}^2$ for your preferred model.

(6)

Explain the economic significance of your preferred model.

(4)

3

Turn over

A2.   **Expenditure on clothing and footwear (£m)   Total consumers' expenditure (£m)**

	current prices	1990 prices	current prices	1990 prices
1982	10925	14447	169372	249852
1983	12120	15441	185611	261200
1984	13168	16261	198820	266486
1985	14912	17615	217485	276742
1986	16646	19169	241554	295622
1987	17848	20204	265290	311234
1988	19023	20780	299449	334591
1989	19847	20662	327363	345406
1990	20876	20876	347527	347527
1991	21412	20817	365469	340037
1992	22097	21455	383490	339652
1993	23528	22665	406569	348164
1994	24838	23854	427394	357845
1995	25899	24852	446169	364046
1996	27434	26516	473509	376648

Source : United Kingdom National Accounts, 1993 and 1997 editions, Tables 4.7 and 4.8.

The above statistics are analysed using the Minitab statistics package, giving the output shown on the following page of this examination paper. Columns 1 to 5 contain the data given in the above table. Column 8, which is not shown on the output, contains the values -7, -6, ..., 6, 7, i.e. it is (column 1) - 1989. (The subcommand **predict** evaluates the fitted regression function for the stated value(s) of the explanatory variable(s) and gives both 95 per cent confidence intervals appropriate to this evaluation.)

What analyses have been carried out?

(10)

Write an economic account of what has been learnt from these analyses.

(10)

(Minitab output on following page)

```
MTB > let c6 = (c2/c3)/(c4/c5)
MTB > let c7 = loge(c6)
MTB > print c1-c7
```

Row	Year	C2	C3	C4	C5	C6	C7
1	1982	10925	14447	169372	249852	1.11554	0.109338
2	1983	12120	15441	185611	261200	1.10458	0.099464
3	1984	13168	16261	198820	266486	1.08539	0.081942
4	1985	14912	17615	217485	276742	1.07721	0.074371
5	1986	16646	19169	241554	295622	1.06275	0.060864
6	1987	17848	20204	265290	311234	1.03638	0.035732
7	1988	19023	20780	299449	334591	1.02288	0.022622
8	1989	19847	20662	327363	345406	1.01350	0.013408
9	1990	20876	20876	347527	347527	1.00000	0.000000
10	1991	21412	20817	365469	340037	0.95701	-0.043945
11	1992	22097	21455	383490	339652	0.91219	-0.091908
12	1993	23528	22665	406569	348164	0.88895	-0.117711
13	1994	24838	23854	427394	357845	0.87181	-0.137184
14	1995	25899	24852	446169	364046	0.85031	-0.162151
15	1996	27434	26516	473509	376648	0.82298	-0.194825

```
MTB > correlate c6 and c8
```

Correlation of C6 and C8 = -0.986

```
MTB > correlate c7 and c8
```

Correlation of C7 and C8 = -0.981

```
MTB >  regress c6 on 1 variable in c8;
SUBC>  note - see text of question regarding the next subcommand;
SUBC>  predict 8.
```

The regression equation is
C6 = 0.988 - 0.0215 C8

Predictor	Coef	Stdev	t-ratio	p
Constant	0.988099	0.004346	227.36	0.000
C8	-0.021528	0.001006	-21.40	0.000

s = 0.01683   R-sq = 97.2%   R-sq(adj) = 97.0%

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	1	0.12976	0.12976	458.01	0.000
Error	13	0.00368	0.00028		
Total	14	0.13345			

Fit	Stdev.Fit	95% C.I.	95% P.I.
0.81588	0.00915	(0.79611,0.83564)	(0.77448,0.85727)

```
MTB > let c9 = c3/20876
MTB > let c10 = c5/347527
MTB > regress c9 on 2 variables in c6 and c10
```

The regression equation is
C9 = 0.987 - 0.738 C6 + 0.777 C10

Predictor	Coef	Stdev	t-ratio	p
Constant	0.9867	0.3917	2.52	0.027
C6	-0.7376	0.2293	-3.22	0.007
C10	0.7773	0.1897	4.10	0.001

s = 0.03685   R-sq = 95.6%   R-sq(adj) = 94.9%

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	2	0.35652	0.17826	131.31	0.000
Error	12	0.01629	0.00136		
Total	14	0.37281			

Turn over

A3. **Gross trading profits (net of stock appreciation) of industrial and commercial companies.**
UK, £m, not seasonally adjusted.

		profits	$\log_e$(profits)	trend	$\log_e$(profits) - trend
1992	Q1	18202	9.809	-	-
	Q2	18261	9.813	-	-
	Q3	17703	9.781	9.850	-0.069
	Q4	21694	9.985	9.864	0.121
1993	Q1	18653	9.834	9.894	-0.060
	Q2	19942	9.901	9.931	-0.030
	Q3	20511	9.929	9.975	-0.046
	Q4	25286	10.138	10.020	0.118
1994	Q1	22763	10.033	10.056	-0.023
	Q2	23261	10.055	10.083	-0.028
	Q3	23449	10.063	10.099	-0.036
	Q4	27540	10.223	10.110	0.113
1995	Q1	23761	10.076	10.128	-0.052
	Q2	24299	10.098	10.150	-0.052
	Q3	26000	10.166	10.176	-0.010
	Q4	29619	10.296	10.209	0.087
1996	Q1	27142	10.209	10.237	-0.028
	Q2	27787	10.232	10.268	-0.036
	Q3	28467	10.257	10.297	-0.040
	Q4	34711	10.455	10.309	0.146
1997	Q1	28996	10.275	10.319	-0.044
	Q2	28830	10.269	10.323	-0.054
	Q3	29550	10.294	-	-
	Q4	34538	10.450	-	-

Source of Profits data : Economic Trends, May 1998, Table 2.11.

The trend shown above is a centred four-quarter moving arithmetic average of the logarithms. Explain what this means showing how the first two values of the trend (i.e. 9.850 and 9.864) given were calculated. What are the advantages of using this method to calculate a trend for such data? (4)

Draw a time chart of profits using a logarithmic vertical axis. (5)

Why is a logarithmic vertical axis often preferred to an ordinary vertical axis? (3)

Use the last column of the table given above to obtain seasonal correction factors for the four quarters of the year and interpret them. Use your factors to seasonally-adjust the data for 1997. (5)

When is this method of seasonal adjustment to be preferred to the simpler differences-from-a-moving-arithmetic-average method? Why? (3)

A4. Purchasers of units in unit trusts hope to benefit from a combination of income paid by their trusts and increases in the value of their units. Some trusts specialise on high incomes, others on potential increases in the value of their units, and a third category seek a compromise between these two objectives. To examine whether the three types of trusts differed in their success, statistics of their total returns (i.e. the growths in their selling prices together with the incomes paid out, as percentages of their selling prices at the beginning of the period) are compiled for the ten years ending 31 December 1996 for random samples of trusts of all three types specialising in investment within the United Kingdom, as follows:

UK Growth trusts:

      121.0 148.4 265.0 49.0 135.2 181.9 169.6   77.0

      150.9 156.8 (Sum = 1454.8, sum of squares = 242706.42)

UK Growth and Income trusts:

      207.9 193.1 203.8 195.8 192.5 225.6 208.2 203.1

      164.3 193.7 (Sum = 1988.0, sum of squares = 397444.74)

UK Equity Income trusts:

      130.1 171.5 132.8 123.9 204.9 216.8 230.9 225.9

      (Sum = 1436.8, sum of squares = 272657.18)

(i) Investigate whether there is substantial evidence that the types of trust varied in their total returns by carrying out an appropriate analysis of variance. Explain the logic of your procedure. On what assumptions is your analysis of variance test based? What corresponding non-parametric test is available? (Do not carry it out.) On what assumptions is your non-parametric test based?

(10)

(ii) Which type of trust had the smallest sample mean?

For that type of trust, use a t-test to test the null hypothesis that the population mean was 200.0%.

Perform a corresponding non-parametric test, stating the null and alternative hypotheses and explaining the logic of your test.

(10)

7

Turn over

BLANK PAGE

SECTION B - ECONOMETRICS

B1.   A random sample of 30 individuals was drawn from a population of commuters who had the opportunity to travel to work by either car or public transport. Sample members were asked which mode of transport they more often employed, and to state their usual journey times for each mode of travel. From this information the following quantities were determined for each sample member:

$$y = \begin{cases} 1 & \text{if travel by car was more often employed} \\ 0 & \text{if travel by public transport was more often employed} \end{cases}$$

$$x = \text{(public transport time - car time)}, \quad \text{in minutes.}$$

The following model was adopted:

$$P(y = 1) = P(Z \le \alpha + \beta x)$$

where Z is a standard Normal random variable, and α and β are fixed parameters. Maximum likelihood estimation yielded:

$$\hat{\alpha} + \hat{\beta} x = -0.060 + 0.030x$$
$$\quad\quad\quad (0.304) \quad (0.013)$$

where figures in brackets are estimated standard errors.

(i)   Provide as full an explanation as possible of the methodology through which the model was estimated.

(5)

(ii)   Provide as full an interpretation as possible of the estimated model.

(5)

(iii) What is the implication of a negative value for the parameter α?

(5)

(iv)   An individual is selected at random who states that it takes 30 minutes longer to travel to work by public transport than by car. Provide an estimate for the probability that this individual will more often travel by car.

(5)

9

B2.   An agricultural economist believes that the mean weekly consumption of beef (y) depends on the price of beef (x_1), the price of pork (x_2), the price of chicken (x_3), and mean income per household (x_4). Prices and mean income are in real terms, i.e. are adjusted for inflation. The following fitted regression was obtained through least squares, based on thirty annual observations:

$$\log \hat{y} = -0.024 - 0.529 \log x_1 + 0.217 \log x_2 + 0.193 \log x_3 + 0.416 \log x_4$$
$$\qquad\qquad (0.168) \qquad\quad (0.103) \qquad\quad (0.106) \qquad\quad (0.163)$$

where figures in brackets are estimated standard errors. The coefficient of determination for the fitted regression was $R^2 = 0.683$.

(i)   Provide as full an interpretation as possible of the fitted model.   (5)

(ii)   Test the null hypothesis that the four variables ($\log x_1$, $\log x_2$, $\log x_3$, $\log x_4$) do not, as a set, have any linear influence on $\log y$. State the assumptions necessary for the validity of your test.   (5)

(iii)   The economist is concerned about the possibility of autocorrelated errors in this model. Why would this possibility cause concern? Outline an appropriate specification test for this possibility.   (5)

(iv)   The economist is also concerned that, over the years, increasing awareness of the effects of high consumption of red meat on health may have influenced demand for beef. If this is the case, how would this influence your view of the estimated regression?   (5)

B3.   (i)   What is meant by saying that the generating process for an economic time series has a unit autoregressive root? Outline a test of the null hypothesis of a unit autoregressive root against the alternative of stationarity.   (10)

(ii)   The time series y_{1t} and y_{2t} are each generated by processes with unit autoregressive roots. What is meant by saying that these series are cointegrated? Derive the consequences for the vector autoregressive representation of a pair of cointegrated series.   (10)

B4.   Write notes on four of the following, including a discussion of their relevance to practical econometric analyses. **(There are 5 marks for each chosen part.)**

(i)   Testing for autocorrelated errors in regressions with lagged dependent variables.

(ii)   Tobit analysis.

(iii)   Two-stage least squares estimation in simultaneous equations systems.

(iv)   The identification problem in simultaneous equations systems.

(v)   Instrumental variables estimation methods.

(vi)   Dummy variables.

SECTION C - OPERATIONAL RESEARCH

C1. (a) Use Fibonacci search to find an interval of length ≤ 0.1 containing the value of x which maximizes the function

$$f(x) = 4 - \left(x + \frac{1}{x} \right)$$

in the interval [0.1, 1.6].

(7)

(b) Under what conditions can we use the Newton-Raphson method for the unconstrained maximisation of a nonlinear function of two variables $f(x,y)$?

Perform one iteration of the Newton-Raphson method to maximise the function
$$f(x, y) = -4x^2 - 3y^2 - 8y^{\frac{1}{2}} + 5x$$

in the upper half-plane (i.e. for $y \geq 0$) with starting point (1,1). Describe in words how you would carry out the next iteration. What evidence do you have that you are getting closer to the maximum value?

(13)

C2. (a) Why are *control variates* used in simulation, and what is the rationale behind their use? Describe how control variates are used, and name two other methods used for this purpose.

(6)

(b) Ten replications of a simulation experiment were performed to estimate the expected time to manufacture a cycle safety helmet. The sample mean time over the 10 replications was 137 minutes, with sample standard deviation 110 minutes. How many more replications are needed in order to estimate the expected time to within ± 20 minutes, with 95% confidence?

(6)

(c) How would you explain, to a non-technical person such as the manager of the cycle helmet factory in (b), the importance of carrying out sufficient replications of a simulation?

(3)

(d) What is meant by *steady-state* in the context of simulation? Describe two methods for determining when steady-state has been reached. Under what circumstances might you wish to collect data from the period before steady-state is attained in a simulation?

(5)

C3. A manufacturing company makes pillow cases, fitted sheets and duvet covers. Each are produced and sold to the retailer in batches of 10. To manufacture these products, two types of skilled labour are required: cutting and machining. The production costs, selling prices and hours of skilled labour required per batch are as follows:

	Pillow cases	Fitted sheets	Duvet covers
Selling price	50	85	105
Production cost	45	65	80
Cutting (hours)	2	1	0
Machining (hours)	0	2	1

Each day, 40 hours of cutting time and 30 hours of machining time are available.

The products are manufactured from stock sheets of material. Three stock sheets are required to produce one batch of pillow cases. However the offcuts from making duvet covers can be used in the production of pillow cases; one batch of duvet covers produces offcuts equivalent to one stock sheet. A maximum of 15 complete stock sheets per day are available for making pillow cases, plus any extra stock sheets created from duvet cover offcuts.

(i)   Formulate the above problem as a linear programme.

(5)

(ii)   Use the simplex method to calculate the daily product mix which will maximise profit. State the total profit and the amount of slack for each constraint. Comment on whether the optimal solution should be followed by the manufacturer.

(8)

(iii) The manufacturer decided to stop producing duvet covers. This releases an extra 45 stock sheets for making pillow cases per day, giving a total of 60 per day. Formulate this new problem as a linear programme, and calculate the new product mix by solving graphically.

(4)

(iv)   Calculate how sensitive the new solution is to a change in the profit on pillow cases.

(3)

C4. (a) Derive the equation

$$p_n = \left(\frac{\lambda_{n-1}}{\mu_n} \right) p_{n-1}$$

for the steady-state probability p_n of n customers in a queueing system, stating the conditions under which it is valid, and explain its importance.

(8)

(b) A market survey firm is paid £1.00 by its clients for each person interviewed. Researchers are sent into a shopping centre where people pass at random at a rate of 60 per hour. It takes a researcher six minutes on average to conduct the interview, and the interview time has an exponential distribution. Nobody queues to be interviewed, but if a researcher is free when someone passes, they will agree to be interviewed. Researchers are paid £7.50 per hour, giving the following financial projection.

Number of researchers (s)	Expected profit per hour (£)
1	1.07
2	1.80
5	0.88
6	-0.90

Calculate the hourly expected profit for $s = 3, 4$ researchers and advise the firm as to how many researchers should be sent out.

(12)

SECTION D – MEDICAL STATISTICS

D1.   A parallel group phase III randomised controlled clinical trial (RCT) is being designed to compare the efficacy of a new anti-hypertensive drug *A* compared to standard drug *B*. The trial protocol is about to be written.

(i)   State the items which should be included in the clinical trial protocol.

(8)

(ii)   In the RCT the primary outcome is the change in blood pressure between baseline and 14 days follow-up measured in mmHg which can be assumed Normally distributed. A difference of δ mmHg in changes in blood pressure between drug *A* and drug *B* is considered clinically important. The common standard deviation of the change in blood pressure between baseline and 14 days is σ. In the trial *n* patients are to receive the new treatment *A* and another *n* are to receive the standard drug *B*.

(a)   Derive an approximate formula for the necessary sample size *n* in terms of type I error (α) and type II error (β), using a two tailed test.

(10)

(b)   The new drug would be considered effective if it reduced mean blood pressure by 10 mmHg more than the standard drug. Evaluate *n* for α (two sided) = 0.05, $\beta = 0.20$ and assuming that the common standard deviation of the change in blood pressure (σ) is 12.5 mmHg.

(2)

D2. Patients were randomised to receive one of two treatments for leg-ulcers, *A* or *B*, and followed up for one year. The healing times, in weeks, for 28 patients in the trial are given below. A * indicates a censored observation.

Treatment *A*:   3, 9, 10, 14, 15*, 15, 17, 20, 27, 28*, 37, 52*, 52*, 52*,
                           52*, 52*

Treatment *B*:   1, 13, 17, 30*, 32*, 46, 48, 52*, 52*,52*, 52*,52*

(i)   Explain what is meant by a censored observation.

(1)

(ii)   Compute Kaplan-Meier healing curves for the two treatment groups and show both on one graph.

(10)

(iii)   Calculate the logrank statistic. Is there a difference in the healing patterns of the two treatment groups?

(8)

(iv)   It is known that age, ulcer duration and ulcer area prior to treatment also influence healing time. What analysis could be used to adjust ulcer healing times for these prognostic variables?

(1)

D3. Describe the direct and indirect methods of age standardising mortality rates, commenting on the differences between these approaches.

(4)

Anderson *et al.* (1985) studied mortality associated with volatile substance abuse (VSA), often called glue sniffing. In this study all known deaths associated with VSA from 1971 to 1983 inclusive were collected. The table shows the age distribution of these deaths for Great Britain and for Scotland, with the corresponding age distributions at the 1981 decennial census. Note that Great Britain includes England, Scotland and Wales.

(i) For Great Britain calculate age specific mortality rates for VSA per year and for the whole period. What is unusual about these age specific mortality rates?

(8)

(ii) Calculate the Standardised Mortality Rate (SMR) for VSA deaths for Scotland.

(4)

(iii) Calculate the 95% confidence interval for this SMR.

(3)

(iv) Does the number of VSA deaths in Scotland appear particularly high?

(1)

Volatile substance abuse mortality and population size, Great Britain and Scotland, 1971-83.

Age group (years)	Great Britain		Scotland	
	VSA deaths	Population (1000s)	VSA deaths	Population (1000s)
0-9	0	6770	0	653
10-14	44	4271	13	425
15-19	150	4467	29	447
20-24	45	3959	9	394
25-29	15	3616	0	342
30-39	8	7408	0	659
40-49	2	6055	0	574
50-59	7	6242	0	579
60+	4	10769	0	962

Anderson et al. 1985, British Medical Journal.

Turn over

D4. (i) Define the sensitivity, specificity, positive predictive value and negative predictive value of a diagnostic test. Derive expressions for the positive and negative predictive values in terms of the sensitivity, specificity and prevalence. Suppose the prevalence of the disease increases while the sensitivity and specificity remain unchanged; explain the effects on the positive and negative predictive values, justifying your answer.

(9)

(ii) The tables below show three artificial sets of test and disease data, based on 130 patients suspected of having a certain disease.

	Disease		
Test 1	Positive	Negative	Total
Positive	5	7	12
Negative	1	117	118
Total	6	124	130

	Disease		
Test 2	Positive	Negative	Total
Positive	4	4	8
Negative	2	120	122
Total	6	124	130

	Disease		
Test 3	Positive	Negative	Total
Positive	3	0	3
Negative	3	124	127
Total	6	124	130

Calculate the sensitivity and specificity for each of the three diagnostic tests.

(3)

Determine the positive and negative predictive values of the tests in populations in which the prevalence of disease is

(a) 5%,

(3)

(b) 10%.

(3)

The three tests were actually based on cut-off values of 1, 2 and 3 units of a biochemical marker. On one graph sketch the ROC curve for these three cut-off values.

(2)

SECTION E – BIOMETRY

E1.   What are the main features of a split-plot design?

(6)

Give an example of a situation where a split-plot design would be preferred to a randomised block design using the same number of plots. Explain why, for your example, the split-plot design would be better.

(4)

In an experiment to test the efficacy of different cultivation methods on the yield of winter wheat, six cultivation methods were assigned to whole plots in each of four replicates, and each plot was split into two sowing dates, early and late sown.

Give the degrees of freedom for each source of variation in the analysis of variance.

(4)

Give the formulae for the standard deviations of the differences between the means of the following:

(i)   Two cultivation methods over both sowing dates.

(ii)   Two cultivation methods at the same sowing date.

(iii)   Two cultivation methods at different sowing dates.

(6)

Turn over

E2.   Define the Logistic Curve and the Gompertz Curve.

(5)

What is the relationship between gradient (dy/dt) and ordinate (y) for each of these curves? Why does that make them suitable for modelling growth of organisms and populations of organisms?

(3)

The following data are taken from standard tables of mean weights in kilograms of female dairy cattle:

Age in months	Jersey females	Ayrshire females
0	24	33
2	41	54
4	72	90
6	110	133
8	147	176
10	178	213
12	204	244
14	230	277
16	253	303
18	273	329
20	291	360
22	310	383
24	332	409
36	388	439
48	408	469

Assuming that the final figure is close to the asymptotic value and the lower asymptote is zero, plot the data and estimate visually the approximate age at which each breed of cattle achieves half its final weight. By considering the symmetry of the plots about the central value, which of the two curves would you expect to give the better fit to these data? Explain your reasoning.

(4)

Bearing in mind the range of values of the data, what relationship might you expect between the variance and the expected weight at any given age? How would this consideration affect the method you would use to fit one of these models to each set of data?

(4)

Assuming that an automatic method for fitting growth curves is available, how would you test the hypothesis that the growth curves for the two breeds of cattle are in direct proportion?

(4)

E3.   Explain the terms *Biological Assay* and *Median Effective Dose*.   (4)

In an experiment to determine the effectiveness of a sulphur compound as a treatment for a fungal disease of potatoes, the following procedure was adopted.

Five concentrations of fungal spores were prepared, and 48 potato leaves from untreated potato plants were inoculated with each concentration of spores, and 48 further leaves from potato plants treated with sulphur were similarly inoculated with each concentration. The numbers of leaves developing disease symptoms were as follows:

Spore concentration per millilitre	Number of leaves diseased Untreated	Number of leaves diseased Sulphur treated
1	9	4
4	20	11
16	40	29
64	46	37
256	48	46

Describe in detail how you would analyse this experiment to measure the benefits, if any, of the sulphur treatment.   (6)

Illustrate your analysis graphically.   (3)

A computer analysis of these data assuming a logistic relationship between the proportion of diseased leaves and the common logarithm (base 10) of spore concentration, and binomially distributed responses, produced the following results:

	Intercept	Slope	Deviance
Separate responses			
Untreated	-1.729	2.700	1.978
Sulphur treated	-2.408	2.169	1.588
Parallel responses			
Untreated	-1.498	2.388	combined
Sulphur treated	-2.654	2.388	5.109

The total deviance for a common model for both treatments was 25.428.

Set out the analysis of deviance, with the correct degrees of freedom, and show how it can be used to test
(i)   whether the logistic model fits the data,
(ii)   whether a common slope may be assumed,
(iii)   whether sulphur treatment reduces disease.   (5)

Estimate the relative effectiveness of sulphur treatment in terms of the difference in median effective spore concentrations for each treatment.   (2)

Turn over

E4.   What are the main principles of experimental design?

(6)

An agricultural field trial to test the effects of the three main plant nutrients, Nitrogen (N), Phosphate (P) and Potassium (K), is to contain all combinations of each nutrient at three equally spaced levels of application, labelled 0, 1 and 2. As a statistician you are advised that the field must be divided into three blocks of nine plots per block. How would you allocate the treatment combinations to blocks so as to ensure that all main effects and two-factor interactions may be estimated?

(4)

Outline the analysis of variance for the above experiment, giving the degrees of freedom for each source of variation.

(4)

How can the degrees of freedom for each main effect and two-factor interaction be partitioned in order to test for linear responses to each nutrient?

(3)

One of the plots is accidentally damaged before harvest and the yield is not recorded. Describe in detail how this will affect your analysis.

(3)

SECTION F – STATISTICS FOR INDUSTRY AND QUALITY IMPROVEMENT

F1.   A small company manufactures springs which are used in vehicle fuel injector systems. The specification for the length of a spring is between 49mm and 51mm. The customer has recently told the company that it must demonstrate that its process capability is greater than 1. You have been asked to provide statistical advice.

(i)   The production manager has taken random samples of 5 springs from each of the last 6 shifts. The results are:

Shift	Sample size	Mean	Standard deviation (divisor (n − 1))
1	5	49.24	0.17
2	5	49.93	0.24
3	5	50.59	0.26
4	5	49.98	0.18
5	5	50.80	0.22
6	5	49.77	0.37

(a)   Give the definition of process capability. Why is it usually thought necessary that it should exceed 1?

(4)

(b)   Calculate the square root of the average of the 6 variances. Why is this a better estimate of the within sample standard deviation than the average of the 6 standard deviations?

(3)

(c)   Considering the 30 spring lengths as one single sample, the mean and standard deviation are 50.05mm and 0.57mm respectively. Bearing this in mind, what advice would you give the company?

(3)

(ii)   A few weeks later you are asked to set up Shewhart mean and range charts for the process. The target value is 50mm and the standard deviation of spring lengths can be assumed to equal 0.30. Set up charts for samples of size 5, showing action and warning lines, and demonstrate their use with the following data. [Lower and upper 0.1% and 2.5% control chart factors for the standard deviation, when setting up a range chart for samples of size 5, are lower: 0.37, 0.85, upper: 4.20, 5.48 respectively.]

Sample number	Sample size	Mean	Range
1	5	50.2	0.80
2	5	49.9	0.26
3	5	49.7	0.73
4	5	49.9	0.52
5	5	50.2	0.48
6	5	49.8	0.35
7	5	50.3	0.92
8	5	50.0	0.53
9	5	50.1	1.07
10	5	49.5	0.85

(10)

Turn over

F2.	A manufacturer of rechargeable batteries is investigating a new product. The critical variable is the duration for which the battery can supply full power. The duration is related to the amount of compound (A) in the mixture surrounding the electrode and the diameter of the electrode (B). An initial experiment gives the following results:

A (g/litre)	B (mm)	Duration (minutes)
20	15	144
24	15	162
20	17	171
24	17	187
22	16	163
22	16	167
22	16	161
22	16	161

(i)	Write down a model for the experimental results, including terms for linear effects and their interaction.

(4)

(ii)	(a)	Estimate the parameters in your model.

(4)

	(b)	Give an estimate of the standard error of the effects.

(5)

(iii)	(a)	Draw an interaction diagram.

(3)

	(b)	Is there evidence of an interaction?

(1)

(iv)	The manufacturer wishes to search for the maximum duration by changing the amount of compound and the electrode diameter. Use an appropriate method to recommend values of A and B to be used in further trials.

(3)

F3. A production process requires at least one of two machines $A1$, $A2$ and then a larger machine B. This is shown schematically below. Both machines of type A are used if they are both working.
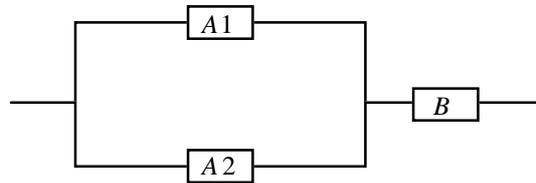
Times between failures are independent. Those for the type A machines have an exponential distribution with mean $1/\lambda_1$, and those for machine B have an exponential distribution with mean $1/\lambda_2$. Repair times are also independent and have exponential distributions with rates ν_1, ν_2 for machines A and machine B respectively. There are several employees who can carry out repairs, so repair work starts immediately and there is no restriction to the number of machines being worked on.

Define six states as follows.

Type A machines	Machine B	State
both working	working	1
one working one failed	working	2
both failed	working	3
both working	failed	4
one working one failed	failed	5
both failed	failed	6

(i) Define a Markov process. What is meant by the statement that a Markov process is ergodic?

(4)

(ii) Draw a state space diagram of the system.

(6)

(iii) Write down the equations you need to solve to find the proportion of time in each state.

(6)

(iv) Assume that

$$\nu_1^2 \nu_2, \; 2\lambda_1 \nu_1 \nu_2, \; \lambda_1^2 \nu_2, \; \lambda_2 \nu_1^2, \; 2\lambda_1 \lambda_2 \nu_1, \; \lambda_1^2 \lambda_2$$

is a solution to the equations in (iii). If λ_1, λ_2, ν_1 and ν_2 are 0.02, 0.01, 1 and 0.5 respectively, for what proportion of the time is the production process feasible?

(4)

F4. An experiment was carried out with the objective of making a spot welding process robust against thickness of metal and level of the welder's experience. A high average strength and a small standard deviation of strength are desirable. The control factors were current, time and type of steel. An L9 orthogonal array was used for the control factors:

Current	Time	Steel
−1	−1	Mild
−1	0	Stainless
−1	1	Galvanised
0	−1	Stainless
0	0	Galvanised
0	1	Mild
1	−1	Galvanised
1	0	Mild
1	1	Stainless

the fourth column being left empty. The noise factors, thin and thick material, and experienced and apprenticed welder, were assigned to the first two columns of an L4 array:

Material	Welder
thin (−1)	experienced (−1)
thin (−1)	apprenticed (+1)
thick (+1)	experienced (−1)
thick (+1)	apprenticed (+1)

The results are summarised below.

Current	Time	Steel	−1 −1	+1 −1	−1 +1	+1 +1 welder material	Mean $\bar{y}$	SD s	S/N
				Strength (kN) y					
−1	−1	M	4.6	5.4	5.6	5.8	5.35	0.53	14.46
−1	0	S	7.8	7.8	8.0	8.4	8.00	0.28	18.05
−1	1	G	3.6	2.9	2.8	1.9	2.80	0.70	8.23
0	−1	S	8.4	9.6	7.0	8.6	8.40	1.07	18.32
0	0	G	6.0	7.1	5.9	5.2	6.05	0.78	15.48
0	1	M	5.6	6.2	8.6	13.1	8.37	3.41	17.14
1	−1	G	6.6	7.8	7.6	6.2	7.05	0.77	16.84
1	0	M	6.2	6.4	6.6	13.6	8.20	3.60	17.06
1	1	S	12.0	12.8	14.6	14.8	13.55	1.37	

(Question F4 continued on next page)

26

(i) *S/N* is the "larger the better" signal to noise ratio defined by

$$S/N = -10 \log_{10}\left(\frac{1}{n}\left(\sum 1/y_i^2\right)\right)$$

where y_i are the four strengths measured for each setting of the control factors. Calculate the missing *S/N* value. Explain why this *S/N* criterion may be appropriate if the objective is to achieve a high mean and a small variance. Plot *S/N* against $\bar{y}$.

(6)

(b) Another heuristically reasonable variable to try and maximise is $\bar{y} - 2s$. The regression of *S/N* on $\bar{y}$ and *s* is:

$$S/N = 7.2 + 1.27\bar{y} - 0.243s$$

with an adjusted R^2 of 88%. Comment on this result.

(2)

(iii) What are the main limitations of this design? Can the design be used to estimate quadratic effects?

(6)

(iv) In the following regressions, x_1 is current; x_2 is time; x_3 is 1 if the steel is stainless and 0 otherwise; and x_4 is 1 if the steel is galvanised and 0 otherwise. Standard errors of coefficients are bracketed underneath.

$$S/N = 16.2 + 2.62x_1 - 0.285x_2 + 3.42x_3 - 2.70x_4$$

$$(0.88) \quad (0.88) \quad (1.76) \quad (1.76)$$

adjusted $R^2 = 68.1\%$

$$S/N = 15.7 + 2.62x_1 - 0.285x_2 + 1.55x_1 \times x_2 + 3.42x_3 - 1.15x_4$$

$$(0.88) \quad (0.88) \quad (1.52) \quad \quad (1.75) \quad (2.32)$$

adjusted $R^2 = 68.5\%$

$$(\bar{y} - 2s) = 2.28 + 0.698x_1 - 0.383x_2 + 5.89x_3 + 1.52x_4$$

$$(0.97) \quad (0.97) \quad (1.95) \quad (1.95)$$

adjusted $R^2 = 45.0\%$

$$(\bar{y} - 2s) = 1.39 + 0.698x_1 - 0.383x_2 + 2.67x_1 \times x_2 + 5.89x_3 + 4.19x_4$$

$$(0.69) \quad (0.69) \quad (1.19) \quad \quad (1.37) \quad (1.81)$$

adjusted $R^2 = 72.7\%$

What tentative conclusions would you draw?

(6)

Turn over

BLANK PAGE