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

GRADUATE DIPLOMA IN STATISTICS, 2003

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.*

The notation log *denotes logarithm to base* ***e****.*
Logarithms to any other base are explicitly identified, e.g. $\log_{10}$.

Note also that $\binom{n}{r}$ *is the same as* ${}^nC_r$.

This examination paper consists of 29 printed pages and 1 blank page (page 30).
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.

THIS SECTION STARTS

ON THE

NEXT PAGE

(page 4)

Turn over

A1. (a) Many economic and other series of monthly, quarterly, etc, data are published in both "not seasonally adjusted" and "seasonally adjusted" form. Explain what is meant by seasonally adjusting a time series and why this is so frequently done. Give examples of such economic time series known to you. [You are NOT asked to explain how such adjustments are made.]

(5)

(b) Quarterly data, not seasonally adjusted, in billions of pounds at current prices, from 1993 first quarter to 2001 second quarter inclusive, relating to United Kingdom exports of services (i.e. receipts from overseas tourists, shipping and travel services, insurance and financial services, etc) are as follows.

1993	Q1	9.2	1997	Q1	13.5
	Q2	9.8		Q2	14.3
	Q3	12.1		Q3	17.1
	Q4	10.3		Q4	14.8
1994	Q1	10.5	1998	Q1	14.5
	Q2	10.7		Q2	15.9
	Q3	13.2		Q3	18.3
	Q4	11.0		Q4	16.0
1995	Q1	11.5	1999	Q1	16.1
	Q2	11.8		Q2	17.6
	Q3	14.3		Q3	19.3
	Q4	12.3		Q4	17.6
1996	Q1	12.5	2000	Q1	17.6
	Q2	13.4		Q2	19.3
	Q3	15.8		Q3	21.5
	Q4	14.2		Q4	19.3
			2001	Q1	18.5
				Q2	19.2

Source: Economic Trends Annual Supplement, 2001 edition, Table 1.18.

They are analysed using the Minitab computer package, as shown in the edited computer output at the end of this question.

(i) What is the relevance of columns C7 and C8 to an analysis of these data? Illustrate your answer by showing how the value 0.1450 in C8 is obtained from the values in C2. (4)

(ii) Obtain seasonal correcting factors for United Kingdom exports of services, and hence calculate seasonally corrected data for the first two quarters of 2001. (6)

(iii) Why were logarithms of the original data obtained at the start of the computer analysis? What would have been the effect of omitting this step in the analysis? (5)

Question A1 is continued on the next page

```
MTB > Print C1

C1
    9.2    9.8   12.1   10.3   10.5   10.7   13.2   11.0   11.5   11.8
   14.3   12.3   12.5   13.4   15.8   14.2   13.5   14.3   17.1   14.8
   14.5   15.9   18.3   16.0   16.1   17.6   19.3   17.6   17.6   19.3
   21.5   19.3   18.5   19.2

MTB > Let C2=log(C1)
MTB > Lag C2 C3 # Note: Col C3 contains C2 values, lagged one quarter
MTB > Lag C3 C4
MTB > Lag C4 C5
MTB > Lag C5 C6
MTB > Let C7 = C2/8  +  C3/4  +  C4/4  +  C5/4  +  C6/8
MTB > Delete 1 2 C7 # Delete first two values in C7 and move others up
MTB > Let C8 = C2 - C7
MTB > Print C1 C2 C7 C8

 Row      C1         C2         C7         C8

   1     9.2    2.21920          *          *
   2     9.8    2.28238          *          *
   3    12.1    2.49321    2.34826     0.1450
   4    10.3    2.33214    2.37576    -0.0436
   5    10.5    2.35138    2.39762    -0.0462
   6    10.7    2.37024    2.41671    -0.0465
   7    13.2    2.58022    2.43630     0.1439
   8    11.0    2.39790    2.45991    -0.0620
   9    11.5    2.44235    2.48215    -0.0398
  10    11.8    2.46810    2.50611    -0.0380
  11    14.3    2.66026    2.53050     0.1298
  12    12.3    2.50960    2.55682    -0.0472
  13    12.5    2.52573    2.58518    -0.0595
  14    13.4    2.59525    2.61560    -0.0203
  15    15.8    2.76001    2.64318     0.1168
  16    14.2    2.65324    2.66092    -0.0077
  17    13.5    2.60269    2.67893    -0.0762
  18    14.3    2.66026    2.69399    -0.0337
  19    17.1    2.83908    2.70810     0.1310
  20    14.8    2.69463    2.73029    -0.0357
  21    14.5    2.67415    2.75202    -0.0779
  22    15.9    2.76632    2.77024    -0.0039
  23    18.3    2.90690    2.79307     0.1138
  24    16.0    2.77259    2.81885    -0.0463
  25    16.1    2.77882    2.83820    -0.0594
  26    17.6    2.86790    2.85677     0.0111
  27    19.3    2.96011    2.87982     0.0803
  28    17.6    2.86790    2.90248    -0.0346
  29    17.6    2.86790    2.92750    -0.0596
  30    19.3    2.96011    2.95251     0.0076
  31    21.5    3.06805    2.97027     0.0978
  32    19.3    2.96011    2.97586    -0.0158
  33    18.5    2.91777
  34    19.2    2.95491
```


Turn over

A2. Changes in inventories are a minor but erratic part of gross capital formation. In order to examine how and whether they can be predicted by gross domestic product (possibly lagged), annual data in millions of pounds at constant prices of 1995 relating to the United Kingdom were collected from Table 13 of *Economic Trends Annual Supplement, 2001 edition,* for the 21 years 1980 to 2000 inclusive.

Changes in inventories were denoted by y and were regarded as endogenous.

Explanatory variables were denoted as follows:

x = gross domestic product
x_{-1} = gross domestic product lagged one year
x_{-2} = gross domestic product lagged two years
$\Delta x = x - x_{-1}$ = increase in gross domestic product over preceding year
$\Delta^2 x = x - x_{-2}$ = increase in gross domestic product over two years
$t = -10, -9, \ldots, 10$, a time trend.

Values of x for 1978 and 1979 were used as necessary to obtain values of these explanatory variables for the entire 21 years considered.

(i) Why should changes in inventories be affected by gross domestic product with any appropriate lags, and/or by changes in it, and/or time? Suggest other possible explanatory variables which might usefully be considered.

(4)

(ii) A regression of y on x and t gave the following results.

$$y = -35059 + 0.05589x - 616t, \qquad R^2 = 0.351, \quad s = 2835 \quad DW = 0.82 \qquad [a]$$
$$\text{(26945)} \quad \text{(0.04162)} \quad \text{(680)} \qquad rss = 144{,}629{,}840$$

(estimated standard errors in parentheses)

Simple correlation coefficients were

$r(yx) = 0.5667, \quad r(yt) = 0.5344, \quad r(xt) = 0.9887.$

(a) Find the partial correlation coefficient $r(xy.t)$.

(b) Test your partial correlation coefficient for statistical significance.

(c) How is your test related to the regression given above?

(6)

Question A2 is continued on the next page

(iii) Further regressions were calculated as follows.

$$y = -1132 + 0.18449x - 0.11157x_{-1} - 0.07535x_{-2} + 106.2t \quad [b]$$
$$(14483)\ (0.03998)\ (0.05538)\ (0.03517)\ (358.5)$$
$$R^2 = 0.903,\quad s = 1163,\quad DW = 2.04,\quad rss = 21{,}637{,}202$$

$$y = -5390 + 0.19341x - 0.12032x_{-1} - 0.06886x_{-2} \quad [c]$$
$$(1736)\ (0.02558)\ (0.04558)\ (0.02675)$$
$$R^2 = 0.902,\quad s = 1131,\quad DW = 2.06,\quad rss = 21{,}755{,}922$$

$$y = -5250 + 0.13401x - 0.12998x_{-2} \quad [d]$$
$$(2002)\ (0.01401)\ (0.01545)$$
$$R^2 = 0.862,\quad s = 1305,\quad DW = 1.72,\quad rss = 30{,}671{,}598$$

$$y = -2920 + 0.13955\Delta^2 x \quad [e]$$
$$(483)\ (0.01338)$$
$$R^2 = 0.851,\quad s = 1320,\quad DW = 1.59,\quad rss = 33{,}117{,}386$$

$$y = -5250 + 0.12998\Delta^2 x + 0.00403x \quad [f]$$
$$(2002)\ (0.01545)\ (0.00336)$$
$$R^2 = 0.862,\quad s = 1305,\quad DW = 1.72,\quad rss = 30{,}671{,}598$$

$$y = -2285 + 0.24348\Delta x \quad [g]$$
$$(543)\ (0.02883)$$
$$R^2 = 0.781,\quad s = 1603,\quad DW = 1.52,\quad rss = 51{,}373{,}392$$

What do you learn from the values of *DW* for the seven regressions given?

(3)

Write an account of the economic conclusions which can be drawn from these analyses, performing such calculations and significance tests as you think desirable.

(7)

A3. The price-earnings ratios of equity shares in three sectors of the United Kingdom stockmarket, as published in *The Times* of 22 May 2002, which may be treated as random samples, were as follows.

Banks	10.2	12.4	15.6	31.1	17.9	15.1	13.5	7.7	15.7	19.0
	16.6	15.6	21.6	17.5						
	$n = 14$		sum = 229.5			sum of squares = 4157.5				
Other Financial	14.9	21.5	17.2	14.1	23.6	16.1	18.2	19.5	20.3	45.3
	20.5	14.8	13.6	18.1	23.6	9.2	20.0	34.3	16.0	29.1
	28.0	10.0	12.7	8.9	10.3	42.4	13.9	22.4	10.8	9.9
	10.0	14.1	5.6	13.4						
	$n = 34$		sum = 622.3			sum of squares = 14091.2				
Insurance	9.1	21.0	9.3	37.3	9.7	35.8	22.6	7.4	7.9	7.7
	17.8	10.9								
	$n = 12$		sum = 196.5			sum of squares = 4500.2				

Note. The price-earnings ratio of a share is its price in the market divided by the profits per share of the company concerned. If the profits per share remain constant, the ratio is the number of years that it would take for the company to earn the price of a share. Any dividend payments to shareholders are irrelevant to the price-earnings ratio.

(i) Why might investors prefer to buy shares with high price-earnings ratios when shares with lower ratios would appear to be more profitable?

(3)

(ii) Perform a one-way analysis of variance on the above data, stating your null and alternative hypotheses. What do you conclude from your results?

(8)

(iii) On what assumptions is your analysis based? How realistic are they?

(1)

(iv) Without performing any tests, consider whether the data suggest possible skewness or the presence of an outlier/outliers. If either or both of these were present, would that invalidate your analysis of variance?

(2)

(v) How could you usefully modify your analysis of variance to take account of (a) skewness, (b) an outlier or outliers?

(2)

(vi) Take the 34 observations relating to Other Financial shares as a random sample. Estimate the population mean with a 95% confidence interval. Are skewness and/or outliers a problem in your analysis?

(3)

(vii) What non-parametric test is available for data such as these? Exactly what does it test? (Do not perform the test.)

(1)

A4. The table shows the resident populations of Scotland and Wales on 30 June 2000, in thousands.

	Scotland		**Wales**	
Age	*Males*	*Females*	*Males*	*Females*
0 – 9	313	298	182	174
10 – 19	331	317	198	190
20 – 44	933	925	496	477
45 – 59	468	486	282	285
60 – 79	385	479	246	285
80 and over	57	128	42	89
	2487	2633	1446	1500

Source: Table 2.2, Monthly Digest of Statistics, May 2002.

(i) Estimate the median male ages in Scotland and in Wales. What assumption(s) is/are necessary in making your estimates? Would it be reasonable to use a similar method to estimate the ninth (top) deciles in the same manner? How else might one proceed to obtain estimates of the ninth deciles?

(4)

(ii) Draw overlapping population pyramids of the above data to compare the age structures in the two countries.

(8)

(iii) What are the most notable differences and similarities that you observe from your pyramids?

(4)

(iv) The means of the four age distributions may be estimated as Scotland Males 37.41, Scotland Females 40.43, Wales Males 38.39 and Wales Females 41.30 years, and the median female ages as 38.96 and 40.23 years respectively. How are the means and medians related to your pyramids?

(4)

SECTION B - ECONOMETRICS

B1. Consider a two-equation simultaneous equation model given by

$$y_{1t} = \beta_{12} y_{2t} + \gamma_{11} x_{1t} + u_{1t}$$
$$y_{2t} = \beta_{21} y_{1t} + \gamma_{22} x_{2t} + \gamma_{23} x_{3t} + u_{2t}$$

Some data were collected on these variables, all of which were measured from their means. The sample sums of squares and cross products were as follows (where, for example, $\mathbf{x_2'y_2} = \Sigma x_{2t} y_{2t} = -9$).

	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{y}_1$	$\mathbf{y}_2$
$\mathbf{x}_1$	5	0	0	20	15
$\mathbf{x}_2$	0	9	0	9	–9
$\mathbf{x}_3$	0	0	6	9	6

(i) Derive the reduced form equations for y_1 and y_2.

(5)

(ii) Show that for the first structural equation it is possible to express the structural parameters in terms of the reduced form parameters in two distinct ways. What are the implications of this for your choice of estimator?

(5)

(iii) Use the data above to estimate the parameters of the reduced form equations.

(6)

(iv) Find both values for the Indirect Least Squares estimates of the parameters of the first structural equation using the result in (ii) above.

(4)

B2. Consider the Friedman permanent consumption-income model

$$y_t^* = \beta_1 + \beta_2 z_t^* \qquad (A)$$

where y_t^* is permanent consumption, z_t^* is permanent income and both are unobservable, and $t = 1, 2, \ldots, T$. Observed consumption and income are denoted by y_t and z_t respectively. Thus $y_t = y_t^* + v_t$ and $z_t = z_t^* + u_t$ where u_t and v_t are random errors with zero means and constant variances σ_u^2 and σ_v^2 respectively. The error terms u_t and v_t are uncorrelated with each other, uncorrelated with y_t^* or z_t^*, and not autocorrelated.

(i) Write model (*A*) in terms of the observable variables, and find the mean and variance of the error term in the regression of y_t on z_t.

(5)

(ii) Find the covariance between z_t and the error term you obtain in part (i) above, and hence prove that the least squares estimator of β_2 is inconsistent.

(5)

(iii) Let x_t represent investment, which is highly correlated with both income and consumption but uncorrelated with u_t and v_t. The method of moments instrumental variable estimators of the two parameters using x_t as an instrument are given by

$$\hat{\beta}_{2(IV)} = \frac{\sum (x_t - \bar{x})(y_t - \bar{y})}{\sum (x_t - \bar{x})(z_t - \bar{z})}$$

and

$$\hat{\beta}_{1(IV)} = \bar{y} - \hat{\beta}_{2(IV)}\bar{z} \, .$$

Show that $\hat{\beta}_{2(IV)}$ is consistent.

(5)

(iv) Suppose $\bar{z} = 230$ and $\bar{y} = 150$.

The following information on the cross products based on 20 yearly observations is available (where, for example, $ZY = \sum (z_t - \bar{z})(y_t - \bar{y}) = 220$).

	X	Z	Y
X	355	365	270
Z		380	220

Use the above information to estimate the parameters β_1 and β_2 using least squares and instrumental variable estimators.

(5)

Turn over

B3. Suppose Y_1, Y_2, Y_3 are independent random variables with identical mean β. Let $\text{Var}(Y_i) = i^2\sigma^2$, for $i = 1, 2, 3$.

(i) Show that the least squares estimator $\hat{\beta} = \frac{1}{3}\sum_{i=1}^{3} Y_i$ is an unbiased estimator of β.

(3)

(ii) Find the variance of $\hat{\beta}$.

(3)

(iii) Consider the weighted estimator $\hat{\beta}_W = \frac{36}{49}Y_1 + \frac{9}{49}Y_2 + \frac{4}{49}Y_3$. Show that $\hat{\beta}_W$ is unbiased and that $\text{Var}\left(\hat{\beta}_W\right) < \text{Var}\left(\hat{\beta}\right)$.

(3)

(iv) Show that $\hat{\beta}_W$ in part (iii) is in fact a generalised least squares (GLS) estimator for the linear model $Y_t = \beta + \varepsilon_t$ with heteroscedastic errors.

(6)

(v) Describe an example of the use of GLS in an empirical econometric investigation known to you.

(5)

B4. Answer **three** of the following. **(There are 6 or 7 marks for each chosen part.)**

(a) What are the econometric implications of cointegration?

(b) What is the difference between stationary, trend stationary and difference stationary processes?

(c) Why is stationarity important in practice?

(d) What is meant by spurious regression?

(e) How would you identify the order of a pure autoregressive process and a pure moving average process?

(f) Explain the Dickey-Fuller test for unit roots. Why is it not possible to use the standard critical values in this test?

SECTION C - OPERATIONAL RESEARCH

C1. A project consists of ten activities whose durations are uncertain. The following precedence table shows the estimated mean and variance of the duration (in days) of each activity.

Activity	*Prerequisites*	*Mean*	*Variance*
A	-	12	2
B	-	20	9
C	-	14	4
D	C	16	16
E	A	28	40
F	B, D	15	4
G	B, D	36	16
H	C	22	7
I	E, F	18	3
J	H	24	11

(i) Draw a network diagram for this project.

(3)

(ii) Assuming the expected activity durations, use an appropriate algorithm to identify the earliest and latest event times if the project is to be completed as soon as possible. Also identify the critical path and its expected completion time.

(4)

(iii) Making reasonable assumptions (which you should state clearly), find the approximate probability that the project will be completed in 75 days or less.

(6)

(iv) Suppose now that the activity durations are **not variable** but are known to be equal to their expected values. However, the activity durations may be reduced by the expenditure of extra money, as shown in the table below. The **maximum possible reduction** denotes the maximum number of days which can be subtracted from the original duration. It is required to reduce the overall project duration to 60 days. Which activities should be modified in order to minimise the total extra cost of achieving the required overall reduction?

(7)

Activity	*Maximum possible reduction*	*Unit reduction cost (per day)*
A	6	100
B	8	800
C	5	1200
D	6	900
E	12	1300
F	10	400
G	16	800
H	15	700
I	7	1000
J	10	1100

C2. (a) Solve the following linear programming problem using the simplex method.

Maximise $6x_1 + 3x_2 + 4x_3$

subject to
$$\begin{aligned} 2x_1 + 2x_2 &\geq 6 \\ 3x_1 + x_3 &= 2 \\ x_1 + x_2 + x_3 &\leq 15 \end{aligned}$$

$$x_1, x_2, x_3 \geq 0$$

State, with reasons, whether your solution is unique.

(10)

(b) A company would like to find the least-cost production schedule for a seasonal product. The demand is 2000 units in May, 4000 in June, 6000 in July, 6000 in August and 2000 in September. The workforce of seasonal workers has to be hired at the beginning of April and kept on until the end of September. These workers need to be trained first; this training costs £200 per worker. The time required in April for training does not affect the production capacity for that month. Each worker can produce 400 units per month on regular time and, if desired, an additional 100 units per month on overtime. The regular wages are £800 per month, and the overtime rate is an additional £400 per month.

Units produced are available for use immediately, but may also be put into storage. The product cannot be kept in storage for longer than two months; for example, if it is produced during April it has to be sold before the end of June. Each unit put into storage incurs a fixed handling cost of £0.50, and the cost of holding one unit in storage for one month is £0.40.

Formulate the problem of finding a production schedule to minimise the total costs as a linear programming problem. **Do not attempt to solve it**.

(10)

C3. (i) Write down the differential-difference equations for the queue size for a single-server queueing system where the arrival rate λ is constant but the service rate μ_n depends on n, the number of customers in the system (including the customer being served). Hence, assuming that every $\mu_n > 0$ and that a steady-state solution exists, derive the steady-state equations for P_n, the probability that there are n customers in the system, in terms of λ, μ_n and P_0. What assumptions are you making about the queueing system in order to derive these equations?

(10)

(ii) Customers arrive at random, at a mean rate of 15 per hour, to buy tickets at a theatre. When there is only one customer in the system, he or she is served by a single server with a mean service time of 4 minutes. However, when there is more than one customer in the system, the server is immediately joined by an assistant who helps to serve the same customer, thus reducing the mean service time to 3 minutes. What is the expected number of customers in the system when it is in steady state?

(10)

[Hint: you may use the identity $\sum_{n=0}^{\infty} nx^n = \frac{x}{(1-x)^2}$ for $|x| < 1$.]

C4. (a) What is meant by sampling from a probability distribution? Explain how you would obtain a random observation from the following discrete probability distribution, given an observation having a (continuous) uniform distribution over the range (0, 1).

Outcome	1	2	3
Probability	0.1	0.7	0.2

If you generated a sequence of random observations from this distribution, how would you modify these values if you discovered that a mistake had been made, an outcome 4 having been overlooked: the chances for outcomes 1 and 2 were as given above, but outcomes 3 and 4 should each have probability 0.1?

(10)

(b) A six-a-side knock-out football tournament involves eight teams, A, B, C, D, E, F, G and H. The first round of matches is A v B, C v D, E v F and G v H. In the next round, the semi-finals, the winner of A v B plays the winner of C v D, and the winner of E v F plays the winner of G v H. The final is (of course) played between the two winning semi-final teams. The tournament thus involves 7 matches in total. The match schedule is shown in the following table.

First round	Semi-finals	Final
Match 1 A v B	*Match 5*	*Match 7*
	Winners Match 1 v	
Match 2 C v D	Winners Match 2	Winners Match 5 v
Match 3 E v F	*Match 6*	Winners Match 6
	Winners Match 3 v	
Match 4 G v H	Winners Match 4	

Unlike real life, all eight teams are identically skilled. In each match, the probabilities of each team scoring 0, 1, 2 or 3 goals (no more are possible!) are as follows.

Goals	0	1	2	3
Probability	0.2	0.4	0.3	0.1

In the event of a draw, a penalty shoot-out takes place, which either side is equally likely to win.

Using the following set of random numbers, simulate the seven matches in this tournament. Firstly, describe how you would simulate a penalty shoot-out, and then describe how you would simulate each match. You should then present your simulation match by match, clearly showing how each simulated score is obtained.

(10)

Random numbers:

0.93	0.76	0.27	0.05	0.07
0.47	0.21	0.87	0.83	0.83
0.86	0.07	0.22	0.87	0.49
0.31	0.74	0.63	0.01	0.42

SECTION D - MEDICAL STATISTICS

D1. A case-control study of a group of peptic ulcer patients and a control group of patients known not to have peptic ulcer was carried out. The control group patients were similar to the ulcer patients with respect to age, sex and socio-economic status. Ulcer patients were classified according to the site of the ulcer, either gastric or duodenal. Aspirin use was ascertained for all subjects.

		Aspirin use		**Total**
		Non-user	*User*	
Gastric ulcer	Cases	39	25	64
	Controls	62	6	68
Duodenal ulcer	Cases	49	8	57
	Controls	53	8	61

Source: Duggan, J.M., et al. Gut 1986.

(i) Briefly explain the difference between a case-control and a cohort study for the evaluation of the association between aspirin use and peptic ulcer. What are the advantages of either relative to the other?

(5)

(ii) (a) Ignoring the site of the ulcer, what would be the odds ratio for the occurrence of ulcer for individuals using aspirin, relative to those not using aspirin?

(2)

(b) Calculate the Mantel-Haenszel estimate of the odds ratio for occurrence of peptic ulcers due to aspirin use, allowing for the site of the ulcer. Calculate a 95% confidence interval for this odds ratio.

(8)

(c) Perform a test of the null hypothesis that aspirin use is unrelated to occurrence of peptic ulcer. Comment on the results of all your calculations.

(5)

D2. A parallel group phase III randomised controlled clinical trial (RCT) is being designed to compare the efficacy of a new serum cholesterol drug *A* with that of standard drug *B*.

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

(8)

(ii) In this phase III RCT, the primary outcome is the change in serum cholesterol between baseline and 28 days follow-up measured in mmol/l, which can be assumed Normally distributed. A difference of δ mmol/l for the change in mean serum cholesterol between drug *A* and drug *B* is considered clinically important. The common standard deviation of the change in serum cholesterol between baseline and 28 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.

(8)

(b) The new drug would be considered effective if it reduced mean serum cholesterol by 0.2 mmol/l more than the standard drug. Evaluate *n* for α (two-sided) = 0.05, $\beta = 0.20$ and assuming the common standard deviation of the change in serum cholesterol (σ) is 0.2 mmol/l. Comment on this estimated sample size, and on what advice you would give the clinician planning the trial.

(4)

D3. (i) Describe the difference between the direct and indirect methods of obtaining age standardised mortality rates. Also discuss when indirect standardisation might be used.

(6)

The United Kingdom Transplant Support Service Authority annually reports the number of solid organ (kidney, heart, liver and pancreas) donors. In 1999, there were 697 organ donors in Great Britain. The table shows the age distribution of these donors for Great Britain and for Scotland, with the corresponding age distributions at the 1991 decennial census. Note that Great Britain consists of England, Wales and Scotland.

Number of cadaveric heart beating solid organ donors by region and donor age, 1 January 1999 to 31 December 1999

Age group (years)	**Great Britain**		**Scotland**	
	Organ donors	*Population*	*Organ donors*	*Population*
0 – 14	54	10,056,261	9	946,010
15 – 29	136	10,092,136	18	1,000,401
30 – 44	159	11,994,092	14	1,187,063
45 – 59	241	9,729,816	25	942,136
60 +	107	10,817,586	8	1,043,590

Source: Transplant Activity 1999. United Kingdom Transplant Support Service Authority, 2000.

(ii) (a) For Great Britain, calculate age specific organ donation rates per million population.

(6)

(b) Calculate the indirect standardised donation rate (SDR) for Scotland.

(4)

(c) Calculate a 95% confidence interval for this indirect SDR.

(3)

(d) Does the number of organ donations in Scotland appear particularly high?

(1)

D4. 623 postnatal women were allocated at random to intervention (311) or control (312) groups in a randomised controlled trial to establish the effectiveness of postnatal support in the community in addition to the usual care provided by community midwives. The intervention consisted of up to 10 home visits of up to three hours duration by a community postnatal support worker in the first postnatal month.

The main outcomes were general health status (including physical functioning) at six weeks post delivery. Physical functioning was assessed by postal follow-up questionnaire issued six weeks postnatally.

(i) Briefly describe the different methods of randomisation that might be used in such a trial.

(4)

(ii) Discuss whether the use of placebos and blinding would be appropriate in such a trial.

(4)

The mean physical functioning scores of the two groups at six weeks postnatally are shown in the table below. A higher score indicates better physical functioning.

		Intervention	*Control*
Physical functioning	*n*	278	265
	Mean	86.9	89.1
	Standard deviation	16.0	15.4

Source: Morrell, C.J., et al. British Medical Journal vol 321 (2000) pages 593–598.

(iii) Suggest reasons why only 543 women provided physical functioning scores at six weeks postnatally compared with the 623 women originally randomised.

(2)

(iv) Do these data suggest that the intervention (of extra postnatal support) alters the physical functioning of new mothers six weeks postnatally? *Stating any assumptions you make*, perform an appropriate hypothesis test to compare the mean physical functioning scores between the intervention and control groups. Comment on the results of this hypothesis test.

(5)

(v) Calculate a 95% confidence interval for the mean difference in physical functioning scores between the intervention and control groups. Discuss whether the confidence interval suggests that women in the intervention group might have a higher level of physical functioning at six weeks postnatally than women in the control group.

(5)

SECTION E - BIOMETRY

E1. An experiment on a growing crop of corn was laid out as three randomised complete blocks. It was part of a programme of experiments whose aim was to find the best combination of fertiliser level and time of application to recommend for future use.

The experiment included four levels, F1 – F4, of a compound fertiliser applied at three times, T1 – T3. The levels and the times were both equally spaced. Crop yield was measured on each plot at the end of the season, and the yields (kg) are summarised in the following table which gives the **totals** of yields of each fertiliser/time combination in the three blocks.

Total yields from three randomised blocks

Fertiliser level		F1	F2	F3	F4	*Time total*
	T1	76	76	104	112	368
Time	T2	75	67	108	100	350
	T3	62	70	114	115	361
Level total		213	213	326	327	1079

Block totals (of 12 plots each) are: I, 354; II, 360; III, 365.
The sum of the squares of all 36 observations is 33951.

(i) Copy and complete the initial randomised block analysis:

Source of variation	Degrees of freedom	Sum of squares	Mean square
Blocks			
Treatments	11	1544.972	
Residual			
Total			

(5)

(ii) Subdivide "Treatments" into Levels, Times and Interaction, and then extract a *linear* component from each effect and a *linear* × *linear* component of interaction.

[NOTE. Orthogonal polynomials for a 3-level factor are L: (–1, 0, 1) and Q: (1, –2, 1); those for a 4-level factor are L: (–3, –1, 1, 3), Q: (1, –1, –1, 1) and C: (–1, 3, –3, 1).]

(7)

(iii) With the aid of a diagram, discuss briefly whether it is useful to write a report that is based mainly on the linear terms found in part (ii).

(4)

(iv) If an experiment of similar size can be carried out on an adjacent site next season, what combinations of F and T would you include? Justify your recommendation.

(4)

E2. The alpha-acid contents (y %) of samples from 18 different hop farms were measured at a central laboratory; the samples may be regarded as a completely random choice from a large number that were available. Five were of hop variety F, seven of N and six of D. Records of mean August temperature (x °C) were available for each of the 18 locations, and it was expected that y would be affected to some extent by x. The data were as follows.

F		N		D	
x	y	x	y	x	y
14.0	4.8	15.3	7.3	16.1	8.1
14.8	5.2	16.3	7.7	16.9	9.3
16.5	6.9	15.1	6.5	14.1	7.0
16.7	7.2	16.6	8.2	15.0	7.8
15.5	6.4	14.5	6.1	14.9	7.3
		16.0	7.3	16.2	8.9
		16.9	8.7		
77.5	30.5	110.7	51.8	93.2	48.4

Some edited extracts from the output of an analysis of covariance program are as follows.

```
SOURCE          DF          SUMSQ(X)     SUMPROD(XY)          SUMSQ(Y)
Varieties                    0.3781          0.4667          10.7694
Residual                    15.1219         13.6867          13.3934
Total           17          15.5000         14.1534          24.1628

REGRESSION OF Y ON X            SOURCE              DF     ADJSSQ(Y)
                                Regression                 12.3876
                                Deviations
                                Residual

LEAST SQUARES MEANS FOR Y                 MEAN(ADJ)      STDEV
                                D           8.157       0.1096
                                F           6.221       0.1202
                                N           7.236       0.1021
```

(i) Plot the data <u>on a single graph</u> using different symbols for the three varieties. Explain briefly why this helps to justify carrying out an analysis of covariance.

(6)

(ii) The acid content will determine which type of beer is to be brewed using which variety, and therefore the main requirement of an analysis is to assess the mean alpha-acid content of each variety. Complete an analysis of covariance to examine the relation between y and temperature x, and produce 95% confidence intervals for the mean alpha-acid content of each variety if grown at the overall mean temperature.

(10)

(iii) Although comparisons between varieties are of less importance, some beers require lower, and some higher, levels of alpha-acid. Comment on whether the results in part (ii) give any help to the brewers in knowing in advance which production line a particular delivery of a single variety should be directed to.

(4)

E3. Observers have reported the numbers of successful nests of a rare bird species in ten breeding areas of roughly equal size as follows.

1 5 1 1 3 2 2 1 3 1

(a) (i) Assuming that the numbers of successful nests form a random sample from a Poisson distribution with mean λ, write down the likelihood function L for these data, and show that

$$\log L = 20 \log \lambda - 10\lambda - 9.7573. \qquad (3)$$

(ii) Write down the maximum likelihood estimate of λ based on these data. (1)

(b) The department's statistician sees these data and asks how many observers actually took part in the survey. Because this information is not available, the statistician advises that, to allow for the likely non-reporting of zero counts, the analysis should be done again, using as a model the truncated Poisson distribution that does not allow zero values. The probability mass function of an observation r from this distribution, with parameter λ, is

$$P(r) = \frac{e^{-\lambda}\lambda^r}{r!\left(1-e^{-\lambda}\right)}, \quad r = 1, 2, 3, \ldots .$$

(i) Show that under this model the reported mean number of successful nests $\bar{r}$ is a biased estimate of λ. (3)

(ii) Write down the likelihood function for the above data under the truncated Poisson model, and obtain an equation which must be solved in order to find the maximum likelihood estimator of λ. Give the key points of an algorithm by which the solution may be obtained. (6)

(iii) Using the table below, plot a graph of the logarithm (log) of the likelihood function under the truncated model, for values of λ given in the table, and use it to estimate the value of the maximum likelihood estimator of λ. Check your estimate by substitution in the equation found in part (ii). (7)

λ	$20 \log \lambda - 10\lambda$	$10 \log (1 - e^{-\lambda})$
1.35	–7.4979	–3.0008
1.40	–7.2706	–2.8315
1.45	–7.0687	–2.6732
1.50	–6.8907	–2.5248
1.55	–6.7349	–2.3857
1.60	–6.5999	–2.2552
1.65	–6.4845	–2.1325
1.70	–6.3874	–2.0173
1.75	–6.3077	–1.9089
1.80	–6.2443	–1.8068
1.85	–6.1963	–1.7107

E4. (i) Define

(a) the *relative potency* of a test compound to a standard compound in biological assay,

(b) the *principle of similarity* as applied to indirect quantitative assays.

(4)

(ii) Discuss briefly how to choose suitable *dose metameters* and *response metameters* in order to use linear regression methods in the analysis of assay data.

[NOTE. A *metameter* is a transformation of the data from the units in which they were originally recorded or expressed.]

(4)

(iii) In the analysis of parallel line assays, list appropriate sets of contrasts to be extracted from the Treatments sum of squares in each of (a) 4-point and (b) 6-point designs. Explain what information each contrast gives, and indicate when a 6-point design would be preferred to a 4-point one.

(6)

(iv) Three possible metameters in a 6-point symmetrical parallel line assay were y, $\log y$ and $1/y$. Extracts from the analyses of variance in terms of each of these are shown in the following table. The residual had 20 degrees of freedom. Discuss which of the three response metameters is the most appropriate.

(6)

Source of variation	*Sum of squares (coded)*		
	y	$\log y$	$1/y$
Between preparations	21.3	15.7	7.0
Regression	103.2	96.5	62.1
Parallelism	25.7	31.8	33.2
Quadratic	44.1	24.6	8.3
Difference between quadratics	8.1	41.7	40.9
Residual	200.0	160.0	80.0

SECTION F - STATISTICS FOR INDUSTRY AND QUALITY IMPROVEMENT

F1. A small company manufactures springs which are used in circuit breakers for computer networks. The specification for the length of a spring is between 49 mm and 51 mm. A customer has recently told the company that it must provide some evidence that its process capability exceeds 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. Their lengths give the following values.

Shift in time order	Sample size	Mean	Standard deviation
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 (C_p). Why is it usually thought necessary that it should exceed 1?

(b) Find a pooled unbiased estimate of the within-samples variance, based on all the data.

Hence estimate the within-samples standard deviation. Comment on any advantages this estimator has over simply taking the mean of the six standard deviations.

(c) Calculate C_p using the estimated within-samples standard deviation.

(d) If the 30 spring lengths are considered as one single sample, the mean and standard deviation are 50.05 and 0.58 respectively. Calculate the process performance index (C_{pk}). What advice would you give the company?

(10)

(ii) A few weeks later you are asked to set up Shewhart mean and range charts for the process. The target value for length is 50 mm. Assume the standard deviation of spring lengths is 0.30 mm. Set up charts for samples of size 5, showing action lines, and demonstrate their use with the following data.

Sample number	Sample size	Mean	Range
1	5	50.2	0.80
2	5	49.9	0.26
3	5	50.1	1.07
4	5	49.5	0.85

(10)

F2. A manufacturer of rechargeable batteries is investigating a new product. The critical response variable is the duration (y hours) for which the battery can supply full power. The duration is related to the concentration (g/litre) of compound A in the material surrounding the electrode and the diameter of the electrode B (mm). The results of two replicates of a 2^2 factorial experiment are given below.

A (g/litre)	B (mm)	Duration (hours)	
20	15	54	56
20	17	52	50
24	15	64	60
24	17	57	57

(i) (a) Calculate the main effect of A, defined as the mean when A is high minus the mean when A is low.

(b) Calculate the main effect of B.

(c) Draw an interaction diagram. Does it suggest that there might be a substantial interaction?

Estimate the interaction effect.

(6)

(ii) The following regression model is fitted to the data:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + e_i \ , \qquad i = 1, 2, \ldots, n,$$

where Y_i is the duration, x_1 is -1 and $+1$ for low and high values of A respectively, x_2 is -1 and $+1$ for low and high values of B respectively, and the e_i are independent Normally distributed errors with mean 0 and standard deviation σ.

(a) Write down the estimates of the coefficients.

(b) Estimate σ.

(c) Calculate a 90% confidence interval for β_3 and comment on it.

(5)

(iii) What values of A and B (in units of g/litre and mm) should a follow-up experiment be centred at if the experimenter wishes to move 2 units (in terms of $\boldsymbol{x} = (x_1, x_2)$) in the direction of steepest ascent up a plane fitted to the data?

(4)

(iv) Explain how the results of this follow-up experiment may be used to decide whether the same regression model is still satisfactory, and to plan further work towards finding settings of x_1 and x_2 that will give an optimal value of Y.

(5)

F3. (a) A system consists of n interlinked components, any of which is in one of two states, either working (1) or failed (0). Let x_i be the state of the ith component. Then $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ represents the states of all the components in the system. The structure function, $\phi(\boldsymbol{x})$, takes the value 1 if, and only if, the system works and the value 0 if, and only if, the system fails.

Let $\phi_s(x_1, x_2, \ldots, x_n)$ and $\phi_p(x_1, x_2, \ldots, x_n)$ represent the structure functions for n components in series and in parallel respectively.

(i) Draw a block structure of a system having the structure function

$$\phi = \phi_s\left(\phi_p\left(\phi_{m1}, \phi_{m2}\right), x_6\right),$$

where $\phi_{m1} = \phi_p\left(x_1, x_2, x_3\right)$ and $\phi_{m2} = \phi_s\left(x_4, x_5\right)$. Write down the structure function for this system in terms of $x_1, x_2, \ldots, x_6$.

(ii) Determine the reliability of the system in part (i) above if: the components fail independently; x_1, x_2 and x_3 have reliability 0.9; and x_4, x_5 and x_6 have reliability 0.98. Give your answer correct to 5 decimal places. What is the weakest feature of the system?

(7)

(b) Twenty electric motors were tested under extreme conditions until they failed, or were withdrawn from the test, or lasted for the duration of the 80-hour test programme. Of the 20 motors, 8 outlasted the test programme. After the test programme had been running for 16 hours, 2 of the motors which were then still working were randomly selected and withdrawn for detailed investigation. The failure times for the 10 motors that failed were 8, 11, 14, 17, 18, 30, 35, 46, 71, 75.

(i) Assume that lifetimes have an exponential distribution. Write down the likelihood function for these data. Estimate the parameter of this distribution. Construct an approximate 90% confidence interval for the mean time to failure, and comment on the validity of the approximation.

(ii) Plot on a graph the observed failure times and the corresponding times expected on the exponential model. Suppose that the scientist running the test wishes to see whether a Weibull distribution might give a better explanation than an exponential. Suggest a suitable graph for doing this, and say how the parameters of a Weibull distribution could be estimated from it. Without doing any more calculation, say what result you would expect for these data.

(13)

F4. A factory has three identical machines which operate independently. The time to failure for any one machine has an exponential distribution with a mean of $1/\lambda$ hours. Working times before different failures are independent random variables.

(i) Suppose that there is one repair crew, that repair times have an exponential distribution with a mean of $1/\theta$ hours and that different repair times are independent random variables.

(a) Find the steady state equations for this situation.

(b) For what proportions of the time in the long run are 0, 1, 2 and 3 machines working?

(c) What is the mean number of machines working in the long run?

(8)

(ii) Suppose instead that there are two repair crews which carry out repairs independently. They do not work together if only one machine is broken down. Now assume that repair times have an exponential distribution with a mean of $2/\theta$ hours.

(a) Find the steady state equations for this situation.

(b) For what proportions of the time in the long run are 0, 1, 2 and 3 machines working?

(c) Will the mean number of machines working in the long run in case (i) be greater or less than in case (ii)? Give a practical reason for your answer. What is the numerical difference in the mean number of machines working if $\lambda = \theta$?

(12)

BLANK PAGE