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

GRADUATE DIPLOMA IN STATISTICS, 2007

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 title.

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 calculators in accordance with the regulations published in the Society's "Guide to Examinations" (document Ex1).

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* ${}^{n}C_{r}$.

This examination paper consists of 27 printed pages, **each printed on one side only**.
This front cover is page 1.
Question 1 of Section A starts on page 2.
There are 24 questions altogether in the paper, 4 in each of the 6 Sections.

SECTION A – STATISTICS FOR ECONOMICS

A1. Statistics on the annual value of goods exported from the United Kingdom in £m at 2003 prices, for each year between 1984 and 2005 inclusive, have been collected from Table 1.3 of United Kingdom National Accounts, 2006 edition, and are denoted by x. Their natural logarithms, correct to two decimal places, are denoted by y, so that $y = \log x$. The variable t, which indexes time, is defined as taking the values –10.5, –9.5, …, 10.5 respectively for these years, so there are 22 pairs of observations (t, y). It is found that $\Sigma t = 0$, $\Sigma t^2 = 885.5$, $\Sigma y = 267.61$, $\Sigma y^2 = 3257.5131$ and $\Sigma ty = 44.575$.

(i) Use ordinary least squares to estimate α and β in the model $y = \alpha + \beta t + u$, where the error terms u are independent N(0, σ^2), and find the standard errors of these coefficients. Also obtain the coefficient of determination (r^2) and estimate σ^2.

(6)

(ii) Test the coefficient of determination for statistical significance. Show mathematically how your test is related to the test of the null hypothesis that $\beta = 0$, which you should also carry out.

(4)

(iii) Use your estimated model to predict y for 2007.

Two different standard errors, and hence two different 95 per cent intervals, are associated with such predictions. Explain the difference between these intervals, and calculate them for your prediction.

(6)

(iv) What value of x for 2007 is predicted by your estimated model? Convert the intervals you have calculated in part (iii) to intervals relating to the prediction of x in 2007.

(4)

A2. The table below gives the operating surplus of UK public non-financial corporations, in £m (not seasonally adjusted) for quarters between 2000 and 2005.

Quarter		Surplus,s	log(s)	trend	log(s) - trend
2000	Q1	10225	9.23	-	-
2000	Q2	9950	9.21	-	-
2000	Q3	10079	9.22	9.21375	0.00625
2000	Q4	9825	9.19	9.21125	-0.02125
2001	Q1	10329	9.24	9.21000	0.03000
2001	Q2	9663	9.18	9.20875	-0.02875
2001	Q3	10278	9.24	9.19625	0.04375
2001	Q4	9477	9.16	9.18125	-0.02125
2002	Q1	9573	9.17	9.15375	0.01625
2002	Q2	9274	9.13	9.13125	-0.00125
2002	Q3	8723	9.07	9.10250	-0.03250
2002	Q4	9414	9.15	9.06750	0.08250
2003	Q1	7707	8.95	9.06375	-0.11375
2003	Q2	8714	9.07	9.07750	-0.00750
2003	Q3	8945	9.10	9.11000	-0.01000
2003	Q4	10197	9.23	9.14375	0.08625
2004	Q1	9247	9.13	9.15125	-0.02125
2004	Q2	9527	9.16	9.14500	0.01500
2004	Q3	8716	9.07	9.12875	-0.05875
2004	Q4	10015	9.21	9.11750	0.09250
2005	Q1	8256	9.02	9.11875	-0.09875
2005	Q2	9702	9.18	9.10750	0.07250
2005	Q3	8617	9.06	-	-
2005	Q4	9250	9.13	-	-

Source: UK Office of National Statistics website, series LRXR

(i) The trend shown above is a centred four-quarter moving arithmetic average of the logarithms. Explain what this means, and show how the first two values of the trend (i.e. 9.21375 and 9.21125) were calculated. Discuss the properties of this method of calculating a trend for such data.

(4)

(ii) Draw a time chart of the logarithm of operating surplus.

(5)

(iii) Why is a logarithmic scale often preferred to an arithmetic scale when plotting time series data?

(3)

(iv) Use the last column of the table given above to obtain seasonal correction factors for the four quarters of the year and interpret these factors. Use your factors to correct the 2005 data for seasonality.

(5)

(v) When is this (multiplicative) method of adjustment by seasonal factors to be preferred to the simpler method of adding (positive or negative) seasonal adjustments to the moving average? Why?

(3)

A3. (a) Give formulae for a Laspeyres (base-weighted) price index in terms of

(i) costs of a fixed basket of commodities,

(ii) weighted averages of price relatives.

[Note that two separate formulae are required, one for (i) and the other for (ii).]

Show that these formulae give the same numerical values.

Give a formula for a Paasche (current-weighted) price index.

Which index is likely to show the larger increase over time? Why?

(8)

(b) [Candidates are advised to relate their answer to part (b) of this question to practical conditions either in the United Kingdom or in their own country.]

It is proposed to compile and publish monthly a price index relating to single parent families. Why is it simpler to compile a Laspeyres than a Paasche index?

(2)

In order to compile this index, data on expenditure have to be collected by interviewing a sample of single parents. In order to take the sample, the country is divided into a number of standard regions, within each of which there are approximately 50–100 administrative or electoral areas. Suppose that a stratified random sample of such areas is taken and interviewing is confined to selected areas.

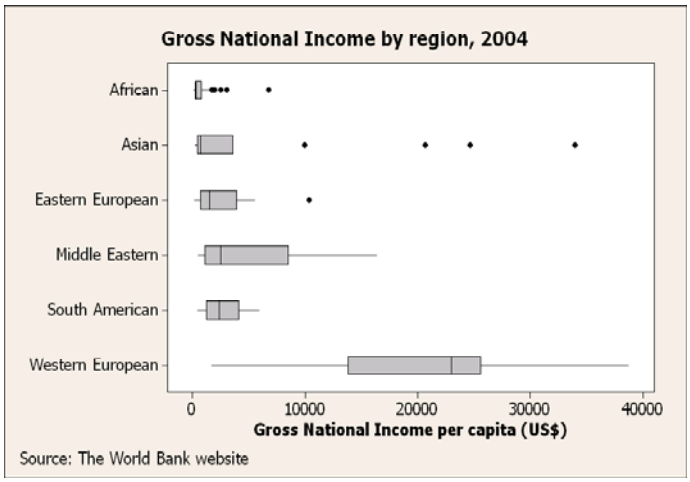
(i) Why should such a sample of areas be taken, rather than interviewing throughout the entire country?

(5)

(ii) Why should the sample of areas be taken using stratification?

(5)

A4. A student has collected data for the gross national income per capita (GNIPC) in 2004 for 134 countries by geographical region. (USA, Canada, Australia and New Zealand are excluded from the data set, as are countries for which GNIPC is unavailable or unreliable.) The data are summarised graphically in the following diagram.



(i) Briefly compare and contrast the six regional distributions on the basis of the graphical evidence.

(3)

(ii) These data yield the summary statistics shown in the following table. The mean and standard deviation, using the $n - 1$ divisor, are in US$ per capita.

Region	N	Mean	StDev
African	42	810	1222
Asian	19	5364	9909
Eastern European	24	2415	2366
Middle Eastern	11	5583	5880
South American	20	2681	1562
Western European	18	21169	9669
ALL	134	5149	8519

The student assumes that, for any country, values of GNIPC in successive years may be modelled as

constant mean value + random error,

where the errors are independent and have constant variance across years and across countries. On this basis, he asks you to perform a one-way analysis of variance using the data for all six regions. Carry out this analysis, and discuss whether it supports the view that income levels per capita vary between countries in different regions.

(6)

Question A4 is continued on the next page

(iii) A further analysis is carried out with the 18 Western European countries excluded. A dummy variable is defined as $x_i = 1$ if country i is either an Asian or Middle Eastern country and as $x_i = 0$ otherwise. Taking the dependent variable y_i as GNIPC, the following regression result is obtained.

$$\hat{y}_i = 1693.4 + 3750.6x_i \qquad r = 0.3391 \qquad n = 116$$
$$\qquad\quad (496.3) \quad (975.9)$$

(The numbers in parentheses are standard errors.)

Does the regression result support the view that, excluding Western European countries, in terms of income per head there are essentially two groups of countries?

(5)

(iv) Write down the standard model and assumptions for one-way analysis of variance, and discuss critically the extent to which the above data (including the Western European countries) conform to these assumptions. Briefly outline how your discussion might be aided if data for two or more years were available.

(6)

SECTION B – ECONOMETRICS

B1. Using conventional notation, the National Income identity may be written as

$$Y = C + I \, .$$

This, together with the consumption function

$$C = \alpha + \beta Y + \varepsilon \, ,$$

where ε is a random variable with zero mean and finite variance, forms a simple macroeconomic model.

(i) Derive the reduced form equations for the model.

(3)

(ii) On what statistical grounds can the use of ordinary least squares to estimate the consumption function be criticised?

(3)

The vectors **Y**, **C** and **I** denote sets of n observations taken on Y, C and I respectively; **1** is the vector whose n elements are all 1; and **X** denotes the $n \times 4$ matrix given by **(1, Y, C, I)**. Annual data (in suitable units) for the United Kingdom for the 14-year period from 1990 to 2003 give the following sums of squares and products matrix.

$$\mathbf{X'X} = \begin{bmatrix} 14 & 9212 & 7290 & 1922 \\ 9212 & 6366300 & 5042293 & 1324007 \\ 7290 & 5042293 & 3994122 & 1048171 \\ 1922 & 1324007 & 1048171 & 275836 \end{bmatrix}$$

(iii) Compute an estimate for β using the method of indirect least squares.

(5)

(iv) Compute an estimate for β using an instrumental variable estimator with I as the instrument for Y. Show that, allowing for rounding errors, your estimate is the same as that obtained in part (iii), and explain why this is so.

(9)

B2. The annual number of housing starts in the United States was modelled by

$$H_t = \alpha + \beta_1 G_t + \beta_2 R_t + \varepsilon_t$$

where, for year t, H_t is the logarithm of the number of housing starts, G_t is the logarithm of the gross national product, R_t is the logarithm of the mortgage interest rate and ε_t is an error term. The model considered for the error terms was $\varepsilon_t = \theta\varepsilon_{t-1} + U_t$, where the $\{U_t\}$ are independently Normally distributed with mean 0 and variance σ_U^2.

Annual data were available for the years 1963–1985, and the model estimated by ordinary least squares was

$$\hat{H}_t = \underset{(1.4)}{-4.759} + \underset{(3.8)}{1.873} G_t - \underset{(4.0)}{1.229} R_t$$

$$n = 23, \quad R^2 = 0.386, \quad Durbin-Watson\ d = 0.794$$

(values in parentheses are standard errors).

(i) Test the hypothesis that there is no first-order serial correlation in the error structure, against the alternative hypothesis most common in economics.

(5)

(ii) Show that the variance σ_ε^2 of the error terms is given by $\sigma_\varepsilon^2 = \dfrac{\sigma_U^2}{1-\theta^2}$, and compute an estimate of θ.

(5)

(iii) In the light of your answer to part (ii), discuss the implications for the validity of using t tests to examine whether the coefficients of the model are zero.

(4)

(iv) Show that an unbiased estimate of the parameter θ may be obtained from a multiple regression of H_t on G_t, R_t, H_{t-1}, G_{t-1} and R_{t-1}. Why might this estimate be unsatisfactory?

(6)

Turn over

B3. (i) Explain what is meant by a *distributed lag model* and an *autoregressive distributed lag model*. Give an example to show how a distributed lag model may arise naturally in an economic context, and explain the terms *short run multiplier* and *long run multiplier* in the context of your example.

(8)

(ii) Briefly discuss the problems that can arise with ordinary least squares (OLS) estimation in distributed lag models.

(4)

(iii) Consider the infinite lag model

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + ... + \beta_k X_{t-k} + ... + u_t \qquad (1)$$

where u_t denotes a zero-mean, constant-variance, serially uncorrelated error term. The Koyck model assumes a geometric dependence of regression coefficients on time lag in the above model, such that $\beta_k = \beta_0 \lambda^k$, k = 1, 2, 3, …, for some λ, $0 < \lambda < 1$. Show that the model (1) may be reformulated as

$$Y_t = \alpha(1-\lambda) + \beta_0 X_t + \lambda Y_{t-1} + v_t,$$

where v_t is an error term which you should specify. Briefly discuss the extent to which the problems of OLS estimation are overcome by the use of the Koyck model.

(8)

B4. Write notes on four of the following. **(There are 5 marks for each chosen part.)**

(a) Misspecification.

(b) Multicollinearity.

(c) Heteroscedasticity.

(d) Assessing forecasting models.

(e) Exogenous dummy variables.

SECTION C – OPERATIONAL RESEARCH

C1. A retailer sells pre-fabricated swimming pools. The expected annual demand for pools is $E(D)$, and demand for pools is uniform throughout the year. (A year can be taken as consisting of 50 retailing weeks.) The cost of placing an order for pools with the wholesaler is K, and delivery of pools from the wholesaler occurs one week after the order is placed. Demands from customers can be backlogged, but at a cost of s per pool short. The cost of holding a pool in stock is h per year.

The retailer's aim is to choose the order size and the re-order point so as to minimise total annual costs. Let q be the optimal order size, let r be the optimal re-order point, and let $1 - F(r)$ be the probability that a shortage occurs during the lead time.

(i) Write down a pair of simultaneous equations for q and $1 - F(r)$ in terms of K, h, s, $E(D)$ and $n(r)$, where $n(r)$ is the expected shortage during the lead time. [**Note.** There is no need to derive these equations.]

(6)

In the rest of this question, suppose that annual demand follows a Normal distribution with mean 250 and variance 800; take K = £2000, h = £5000 and s = £5000.

(ii) Use the assumption that the expected demand during the lead time is small to obtain an approximate value for q.

(2)

(iii) Let X be the demand during the lead time. Using a Normal distribution for X and the value for q obtained above, calculate r.

(6)

(iv) Find the level of safety stock and the maximum expected level of inventory, using the optimum inventory policy.

(2)

(v) Suppose that the lead time is no longer fixed but instead is random. How would this affect the level of safety stock? Briefly explain your answer.

(4)

C2. (a) The Pareto distribution with parameters a, b ($a > 0$, $b > 0$) has the density

$$f_{a,b}(x) = \begin{cases} C_{a,b} x^{-1-a} & \text{for } x \geq b, \\ 0 & \text{otherwise.} \end{cases}$$

(i) Find the constant $C_{a,b}$ and the corresponding distribution function $F_{a,b}(x)$.

(3)

(ii) Using the inversion method obtain an explicit algorithm for generating a random observation from this distribution.

(5)

(iii) Comment briefly on why the rejection method would be unsuitable for simulating Pareto random variables.

(2)

(b) The following is a sample of independent $U(0, 1)$ observations.

0.6154 0.5832 0.8501 0.7700

Let $X(t)$ be the number of customers in the $M/M/1$ Markovian queueing process with $X(0) = 2$. Take the arrival rate as constant, $\lambda = 1$, and the service rate as $\mu(j) = j$ for $j = 1, 2, 3, \ldots$ when $X(t) = j$. Use the given $U(0, 1)$ observations to simulate the times and natures of the first two changes in $X(t)$; explain your methods clearly.

(10)

Turn over

C3. Consider the linear programming problem whose first and final simplex tableaux follow. Note that s_1 and s_2 are slack variables and z is the objective.

BV	x_1	x_2	x_3	s_1	s_2	RHS
s_1	1	1	3	1	0	40
s_2	2	1	3	0	1	60
z	–4	–3	–1	0	0	0

BV	x_1	x_2	x_3	s_1	s_2	RHS
x_2	0	1	3	2	–1	20
x_1	1	0	0	–1	1	20
z	0	0	8	2	1	140

(i) Write down the linear programming problem described by the first tableau and then use the final tableau to determine the optimal solution and optimal value of the objective function.

(2)

(ii) Write down the dual problem and give its optimal solution and the optimal value of its objective function.

(2)

(iii) In the original problem, let c_i be the coefficient of x_i in the objective function, and let b_j be the RHS of the jth constraint. For each of c_1, c_3, and b_1 separately, determine the range of values for which the final basis does not change.

(10)

(iv) What will be the new basis if the value of c_3 is increased to slightly above the upper bound found in (iii)?

(3)

(v) State the *Complementary Slackness Theorem* and demonstrate that it holds for this problem.

(3)

C4. The following table describes a project management problem.

Job	*Duration (days)*	*Immediate Predecessors*
A	3	—
B	4	—
C	5	A
D	2	A
E	4	B
F	3	B
G	3	D, E
H	5	D, E
I	4	C, G
J	5	H, F

(i) Draw the network representing this problem.

(4)

(ii) Compute the total float and free float for each job.

(5)

(iii) Show that there is a unique critical path, and identify it.

(2)

(iv) Draw the Gantt Chart for this problem.

(4)

(v) Re-draw the network for the case where G is also an immediate predecessor of J. Does the critical path change?

(5)

SECTION D – MEDICAL STATISTICS

D1. (i) Explain the difference between a matched and an unmatched case-control study. (2)

We are interested in determining the effect of helmets on the risk of head injuries in skiers and snowboarders.

(ii) Explain why it is not possible to estimate the relative risk directly in a case-control study. (2)

(iii) Explain what is meant by the *odds* of an event. What is the *odds ratio* for exposure and risk? (2)

The data in the table below describe an unmatched case-control study to determine the effect of helmets on the risk of head injuries in skiers and snowboarders. The 693 cases (skiers or snowboarders with head injuries as reported by the ski patrols) and 3294 controls (skiers or snowboarders with non-head injuries who were reported by the ski patrols at the same ski area as the cases) were cross-classified according to helmet use and sex.

Results of an unmatched case-control study of the effectiveness of helmet use in skiers and snowboarders

Males	**Helmet use**		
	Yes	*No*	*Total*
Cases with head injury	158	263	421
Controls	655	802	1457
Total	813	1065	1878

Females	**Helmet use**		
	Yes	*No*	*Total*
Cases with head injury	111	161	272
Controls	788	1049	1837
Total	899	1210	2109

*Source: B. E. Hagel et al. BMJ **330** (2005), 281.*

(iv) Combining the frequencies for males and females, what is the odds ratio for the occurrence of head injuries for helmet users, relative to non-helmet users? Comment on the result. (2)

(v) Calculate the Mantel-Haenszel estimate of this odds ratio, allowing for sex. Calculate an approximate 95% confidence interval for this odds ratio. (8)

(vi) Perform a test of the null hypothesis that helmet use is unrelated to occurrence of head injuries. Comment on the results of all your calculations. (4)

D2. Below are the results of a randomised controlled clinical trial to evaluate an occupational therapy intervention which aims to improve outdoor mobility after stroke. The Intervention group received up to seven visits by an occupational therapist and was compared with a Control group who received only the standard care. The primary outcome measure (Success) was the response to the question whether the participant got out of the house as much as he or she would like four months after the trial began.

		Success		
		Yes	*No*	*Total*
Treatment	*Intervention group*	56	26	82
	Control group	30	56	86

(i) Calculate the proportions in the two groups having a successful outcome after four months. Stating any assumptions that you make, perform an appropriate hypothesis test to compare the proportions of patients in the Intervention and Control groups who got out of the house as much as they wanted. Comment on the results of this hypothesis test, and whether the use of a continuity correction would make any difference to your inference. (5)

(ii) Calculate an approximate 95% confidence interval (CI) for the difference between the two groups in the proportion of patients who got out of the house as much as they wanted. Does the CI estimated from these data suggest that patients in the Intervention group have a better outcome at four months than patients in the Control group? State the assumptions required for your calculation to be valid, and comment on whether the CI gives any additional information to that obtained in part (i). (5)

A secondary outcome measured was the number of outdoor journeys made in the past month.

	Intervention Group (n = 82)		**Control Group (n = 86)**	
	Mean	*Standard deviation*	*Mean*	*Standard deviation*
Outdoor journeys in the past month	37.0	32.6	14.0	21.5

(iii) Perform an appropriate hypothesis test to compare the mean number of outdoor journeys in the past month between the Intervention and Control groups. Calculate an approximate 95% confidence interval (CI) for the difference in the mean number of outdoor journeys in the past month between the Intervention and Control groups. Comment on the results of this hypothesis test and confidence interval.

Discuss carefully the assumptions necessary for your calculations, and whether they appear to be justified for these data. Explain whether or not you would make the same assumptions if the sample sizes had been much smaller, but the values of the means and standard deviations were the same as those above. (10)

*Source of data: P. A. Logan et al. BMJ **329** (2004), 1372–1375.*

D3. Describe the direct and indirect methods of age standardising mortality rates, commenting on the differences between these approaches and the situations in which an indirect method should be used.

(5)

The numbers of heart-beating cadaver donors for 2000, broken down into 10-year age bands, were obtained from United Kingdom Transplant (UKT) data, and are shown in the table below. Organs have to be transplanted very soon after someone has died and they can only be donated by someone who has died in hospital. Usually organs come from people who are certified dead while on a ventilator in a hospital intensive care unit, generally as a result of a major accident like a car crash, a brain haemorrhage or stroke. The table shows the age distribution of these donors for the United Kingdom as a whole and for Scotland separately, with the corresponding population age distributions at the 1991 decennial census. Note that the United Kingdom includes Scotland.

(i) For the United Kingdom, calculate the age-specific organ donation rates for the whole period. Point out any unusual features about these age-specific organ donation rates.

(8)

(ii) Use an indirect method to calculate the Standardised Donation Rate (SDR) for organ donations for Scotland.

(4)

(iii) Calculate an approximate 95% confidence interval for this SDR.

(2)

(iv) Comment on whether the number of organ donations in Scotland appears particularly high compared with the UK as a whole.

(1)

The number of heart-beating cadaver donors for 2000 and population size, UK and Scotland (UKT, 2000)

	UK		**Scotland**	
Age band	*Donors*	*Population*	*Donors*	*Population*
0–9	12	7395033	1	610385
10–19	78	7524570	8	647952
20–29	83	7677822	11	662645
30–39	90	9524953	6	818884
40–49	165	7895663	12	707375
50–59	206	7262395	17	620934
60–69	90	5429785	6	493845
70–79	14	4293963	1	369438
80+	0	2384904	0	183142

D4. (a) Briefly describe the scope of Phase I, II, III and IV clinical trials.

(5)

(b) A parallel group phase III multi-centre randomised controlled clinical trial (RCT) protocol is being designed to compare the efficacy of two creams for the treatment of eczema (a skin disease). The trial will compare a new cream A with the standard cream B, and the primary outcome is whether or not a patient's eczema has healed after 28 days of treatment. The proportions of patients whose eczema heals after 28 days on the new treatment (cream A) and standard treatment (cream B) are denoted by p_a and p_b respectively. In the trial n patients are to receive the new treatment A and another n are to receive the standard treatment B. The null hypothesis is that $p_a = p_b = (P_1 + P_2)/2$; this is to be tested against the alternative hypothesis $p_a = P_1$, $p_b = P_2$, where P_1 and P_2 are specified values and $P_1 \neq P_2$.

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

(8)

(ii) The eczema healing rate of patients after 28 days of treatment with the standard cream B is approximately 50%. The new cream A would be considered effective if it increased the eczema healing rate to 60%. Evaluate n for α (two sided) = 0.05, $\beta = 0.20$.

(2)

(c) Describe the different methods of randomisation that might be used in a multi-centre Phase III clinical trial of the type described in part (b).

(5)

SECTION E – BIOMETRY

E1. Two factors, A and B, are each used at three levels 0, 1 and 2 in an experiment. The experimenter considers that these levels may be regarded as "equally spaced" for the purpose of this study. Factor A is the background noise level at which people carry out a task in a commercial packing shed, and factor B is the complexity of the task they have to do. Code 0 for A is the lowest noise level, and for B it is the simplest of the three tasks used in the study. The quality of the task when finished is assessed, and the time taken is recorded. These are combined into a "performance score" y.

Each of the nine combinations of levels of A and B is used twice, the people taking part are all familiar with the three tasks in factor B, and the allocation of the nine combinations to the 18 people taking part is random. The scores y are listed in the following table, which has been simplified by subtracting the minimum observed score from all of them. The experimenter considers that she can assume these data have an underlying Normal distribution. The bold figures are the totals of the two replicates of each treatment combination.

		Level of B		
		b_0	b_1	b_2
Level of A	a_0	55, 54 **109**	46, 38 **84**	33, 46 **79**
	a_1	45, 52 **97**	39, 42 **81**	15, 11 **26**
	a_2	44, 38 **82**	18, 23 **41**	2, 0 **2**

$\Sigma y = 601$. $\Sigma y^2 = 25303$. Totals for A are 272, 204, 125. Totals for B are 288, 206, 107.

(i) The sum of squares for the 9 treatment combinations is to be subdivided as shown in part (ii). Write down a table showing the coefficients required to construct each of these 8 contrasts among the treatment totals. (6)

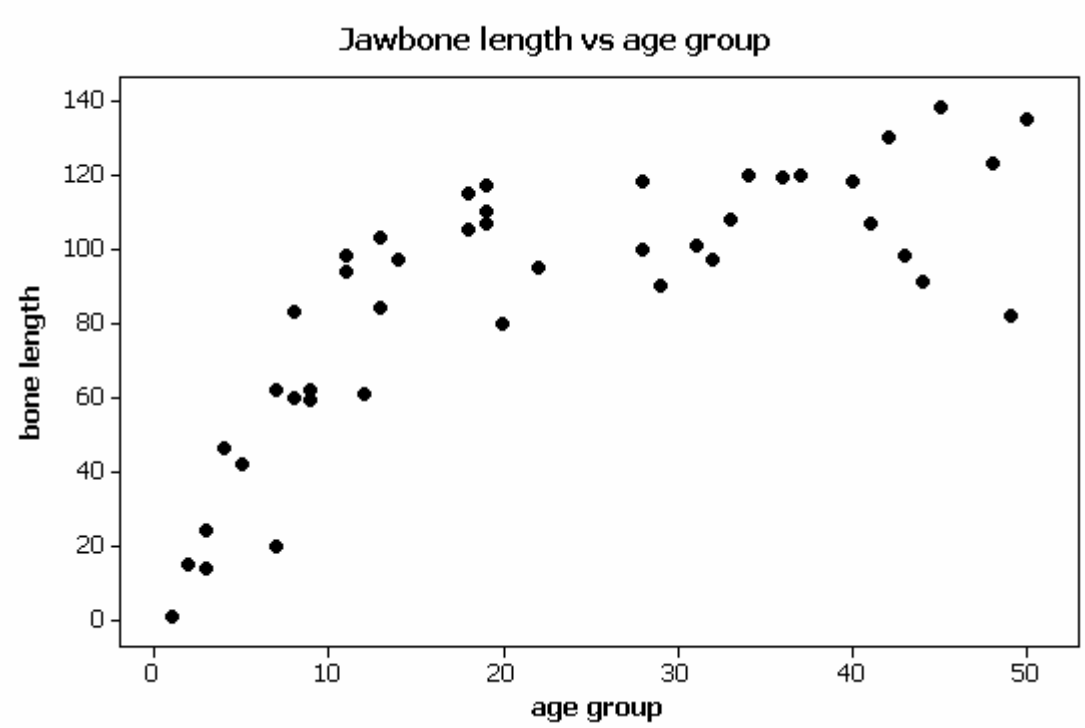
(ii) Copy and complete the following analysis of variance, and carry out appropriate F tests. (7)

Source of variation	Degrees of Freedom	Sum of Squares
Linear A		1800.75
Quadratic A		
Linear B		
Quadratic B		8.03
Linear A × Linear B		
Linear A × Quadratic B		13.50
Quadratic A × Linear B		42.67
Quadratic A × Quadratic B		
Residual		186.50
Total		

(iii) With the aid of the statistical tests and a suitable graph, write a report on the results of the analysis, commenting in particular on any unusual features. Recommend treatment combinations to be used in a further study on this topic. (7)

E2. The scatter diagram below shows 43 pairs of data on the jawbone length (y mm) in deer of various age-groups (x). The ages are grouped into three intervals per year, according to the times at which the measurements became available, and are numbered 1–50.

It is required to find a statistical model which explains the data satisfactorily.



The equation $y = \alpha - \beta\exp(-\gamma x)$ is proposed, where $\alpha, \beta, \gamma > 0$.

(i) Find the normal equations (built up in the usual way by the mathematical least-squares minimisation method) for estimating α, β and γ. Comment briefly on the difficulties that would arise in attempting to solve them.

(4)

(ii) The proposed equation can be written in the form $(\alpha - y) = \beta\exp(-\gamma x)$. This can then be simplified by taking logarithms. Explain what further information is needed to carry out the required estimation. With reference to the scatter diagram, and by considering the properties of the curve originally proposed as a model, make rough estimates of α, β and γ.

[You may use the result that a linear regression of y on x for the animals in age-groups up to 10 is $y = -0.243 + 7.438x$, standard errors of the intercept and slope respectively being 9.58 and 1.56, with $R^2 = 67.1\%$.]

(5)

Question E2 is continued on the next page

(iii) A statistician (A) decides to use an initial estimate of α as 140, and fits the simplified form of part (ii), obtaining the following estimates of intercept and slope. For intercept, the estimate is 4.618 with standard error 0.159, and the slope is –0.0408 with standard error 0.0059. The value of R^2 is 54.1%.

A colleague (B) comments that the data seem very variable in the upper age-groups, and it might be better to set the initial estimate of α at 125 and delete the three data pairs for which y was above this value. After doing this, the estimate of intercept is 4.335 with standard error 0.205 and the slope is –0.0487 with standard error 0.0082. The value of R^2 is 48.3%.

Finally another colleague (C) says that in view of the greater life-expectation of females in this species it would be a good idea to omit the last 9 items in the age-groups 40+ and to set the initial estimate of α at 125. This leads to an estimate of intercept equal to 4.60 with standard error 0.172 and of slope equal to –0.0702 with standard error 0.0087. The value of R^2 is 67.1%.

Find the estimates of β and γ that would have been obtained by A, B and C.

(6)

(iv) In reviewing all these attempts, you note that the following large standardised residuals have been flagged.

- In A's attempt, there were large residuals at the values $x = 45$ and $x = 49$.
- In B's attempt, there was again a large residual at $x = 49$.
- In C's attempt, there was a large residual at $x = 20$.

State and discuss the assumptions needed in the regression calculations that have been used, and comment on the results of all these attempts to find a satisfactory model.

(5)

E3. (a) (i) A response Y is calculated as the ratio of two measurements, $Y = A/B$. The mean and variance of A are (μ_1, σ_1^2) and those of B are (μ_2, σ_2^2). The correlation coefficient between A and B is ρ. Using a rapid approximate first-order method (the delta method), show that the variance of Y is

$$\mathrm{Var}(Y) = \left\{\sigma_1^2\mu_2^2 + \sigma_2^2\mu_1^2 - 2\rho\mu_1\mu_2\sigma_1\sigma_2\right\} / \mu_2^4 .$$

(4)

(ii) For Normally distributed measurements A, B, Fieller's Theorem can be used to obtain confidence intervals for Y. [*You may assume the result of Fieller's Theorem without proof.*] Under what circumstances will limits obtained using the approximation in part (i) be very close to those that would be found using Fieller's Theorem?

(2)

(b) (i) Explain the concept of a *tolerance distribution* in biological assay.

(3)

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

(2)

(iii) Define the *relative potency* of a test compound to a standard compound in biological assay.

(2)

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

(7)

E4. (a) In inverse binomial (sequential) random sampling to estimate the proportion p of members of a population having a certain characteristic, a quota of r members (where $r \geq 2$) of this type is set. The sample items are selected for examination one at a time at random, and sampling stops as soon as r members with the characteristic have been found. Let the random variable N denote the number of items sampled to achieve this.

Show that the probability mass function of N is

$$P(N = n) = \binom{n-1}{r-1} p^r (1-p)^{n-r} \qquad \text{for } n = r, r+1, r+2, \ldots .$$

Show that $E(N) = r/p$. Show also that $\hat{p} = (r-1)/(N-1)$ is an unbiased estimator of p.

Assuming that an unbiased estimator of $\text{Var}(\hat{p})$ is $\hat{p}(1-\hat{p})/(N-2)$, find the precision of the estimator $\hat{p}$ relative to the estimator obtained by ordinary binomial sampling in which r of special type have been found in a random sample of fixed size n.

State, with reasons, the circumstances under which the inverse binomial method is worth considering.

(7)

(b) A field of sugar beet was sampled to estimate the mean percentage sugar content of the crop. The field was divided into a large number of plots of a standard size; 20 of these were selected at random and in each selected plot 5 plants were selected at random. The percentage sugar content of each plant was determined. The corrected sum of squares between plots was 84.74 and that between plants within plots was 176.80.

(i) Estimate the variance components for these two sources of variation, ignoring finite population corrections and stating clearly any other assumptions that must be made.

(4)

(ii) Comment on the results obtained in (i), write down the formula for the standard error of the estimate of the field mean in terms of number of plots sampled (n) and number of plants taken per plot (r), and calculate this standard error for the scheme used.

(3)

(iii) The process is to be repeated in the following season. Suppose that on this occasion the standard error of the estimate of the field mean is required to be no more than 0.1. Also it costs three times as much to locate each plot as it does to sample a plant, so that the total cost is $\$(3 + r)n$. Examine how many plots will be required, and how the total costs compare, in each of the cases $r = 10, 5, 4, 3$, assuming that the variance components are the same as those found above. Comment on the results.

(6)

SECTION F – STATISTICS FOR INDUSTRY AND QUALITY IMPROVEMENT

F1. An electronics company manufactures a small circuit breaker. The specification is that the open-time should be less than 50 micro-seconds. The standard deviation of open-times when the process is running satisfactorily is 4 micro-seconds. The target mean open-time for the manufacturing process is 35 micro-seconds.

(i) Calculate a suitable process capability index if the process is on target. If the process is on target, and open-times have a Normal distribution, estimate the expected number of circuit breakers per million that will be outside the specification.

(3)

(ii) Each day, a random sample of four circuit breakers is drawn, and the open-times in micro-seconds are measured. Results for three days are given below. Set up Shewhart mean and range charts and demonstrate their use with these data.

(6)

Day	*Open-times*				*Mean*	*Range*
1	29.0	36.2	37.8	36.7	34.93	8.8
2	33.4	39.1	30.7	35.4	34.65	8.4
3	38.9	29.1	37.0	42.0	36.75	12.9

[You are given that the factors for the process standard deviation that give lower and upper 0.1% points for the range of a random sample of size 4 from a Normal distribution are 0.20 and 5.31 respectively.]

(iii) Suppose now that the process mean increases to 40 micro-seconds. Calculate the following quantities.

(a) The process performance index. (2)

(b) The expected number of circuit breakers per million that will be outside the specification. (2)

(c) The probability that the next sample mean falls above the upper action line in your Shewhart mean chart. (2)

(iv) The following open-times in micro-seconds were measured on ten consecutive circuit breakers taken from the production line: 41.1, 44.3, 40.6, 36.7, 34.3, 34.2, 32.5, 31.8, 32.1, 32.4. Calculate the lag 1 autocovariance and autocorrelation. Comment on these results.

(5)

[You are given that the mean of the data is 36.0 and the sum of squared deviations from the mean is 180.74.]

F2. An experiment was conducted to investigate the transfer of heat from molten glass to the mould during the manufacture of bulbs for TV screens. Three factors were considered: glass temperature (*A*); cooling time (*B*); and glass type (*C*). Each factor was set at either a low or a high level and coded –1 and +1 respectively. A 2^3 factorial design was performed twice with a random order for the 16 runs. The response was the mould temperature (*y*, degrees Celsius). The results are summarised below in standard order.

A	*B*	*C*	*Mould temperatures*	*Mean mould temperature*
–1	–1	–1	459, 473	466
1	–1	–1	469, 475	472
–1	1	–1	471, 461	466
1	1	–1	470, 478	474
–1	–1	1	454, 458	456
1	–1	1	460, 466	463
–1	1	1	483, 471	477
1	1	1	481, 475	478

The overall mean is 469.0.

(i) A regression model that allows for the three main effects and the interaction between *B* and *C* is fitted. The model has the form

$$Y_i = \beta_0 + \beta_1 A + \beta_2 B + \beta_3 C + \beta_4 INT + E_i ,$$

where *INT* is the interaction term and the E_i are independent $\mathrm{N}(0, \sigma^2)$ random errors.

(a) Obtain the least squares estimates of β_0 and β_1.

(3)

(b) Express the interaction in terms of the factors *B* and *C*, and calculate the least squares estimate of its coefficient.

(3)

(ii) Suppose now that the least squares estimates of β_2 and β_3 are 4.75 and –0.50 respectively. Calculate the residuals when *A*, *B* and *C* are all at the low level.

(2)

(iii) The residual mean square from the regression model is 31.18. Calculate a 90% confidence interval for the coefficient of the interaction.

(4)

Question F2 is continued on the next page

(iv) Predict the mean value of mould temperature for glass type +1, when glass temperature and cooling temperature are set at the mid-points of their ranges. By ignoring uncertainty in the estimates of the parameters of the regression model, give very approximate 95% limits of prediction for the mould temperature of one run of the process at these conditions. Would you expect these limits of prediction to be too narrow or too wide?

(3)

(v) The experimenter wishes to get a more accurate assessment of the combined effects of A, B and C on y, with the range for A, B and C extended to –1.5 to 1.5 in coded units. What further experimental runs would you make? How many degrees of freedom would you have for estimating the standard deviation of the errors?

(5)

F3. Ten randomly selected DC electric motors for use in robotic planet explorers were subjected to a test under extreme conditions for 500 hours. Eight failed after 71, 147, 189, 197, 216, 312, 332 and 353 hours respectively. Two lasted longer than 500 hours.

(i) Assume that lifetimes have a Weibull distribution with cumulative distribution function (cdf)

$$F(t) = 1 - \exp(-\lambda t^{\alpha}), \qquad t \geq 0,$$

where λ and α are positive parameters.

(a) Show that

$$\log t = \frac{-\log \lambda}{\alpha} + \frac{\log(-\log(1-F))}{\alpha}. \qquad (2)$$

(b) Estimate the parameters of this distribution by eye, using a graphical method based on the approximate relationship

$$F(E[T_{i:n}]) \approx \frac{i}{n+1},$$

where $T_{i:n}$ denotes the *i*th smallest time to failure in a sample of size *n*. (6)

(c) Hence estimate the probability that a randomly selected motor subjected to these conditions fails before 48 hours. (1)

(ii) Assume now that lifetimes have an exponential distribution with rate $\lambda > 0$.

(a) How is the exponential distribution related to the Weibull distribution in (i)? (1)

(b) Find the maximum likelihood estimate of λ for the data above. (4)

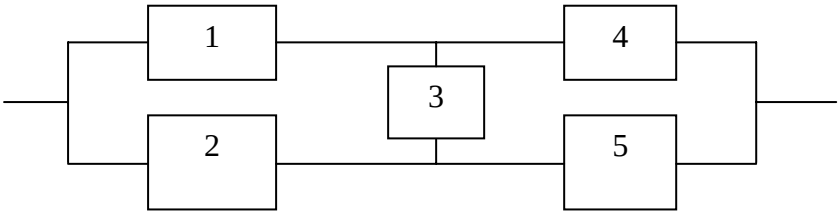
(c) Hence estimate the probability that a randomly selected motor subjected to these conditions fails before 48 hours. (1)

(iii) (a) Derive the hazard function for the Weibull distribution. (2)

(b) On a single sketch, show the hazard functions for the distributions you fitted in parts (i) and (ii). (2)

(c) In relation to this sketch, comment on the difference between your estimates of the probability that a randomly selected motor fails before 48 hours. (1)

F4. The system shown below works if there is a path from left to right.



(i) Write down the structure function for the system in the form

$$\phi(x) = x_3\phi_1(\tilde{x}) + (1-x_3)\phi_2(\tilde{x}),$$

where $x = (x_1, x_2, x_3, x_4, x_5)$ and $\tilde{x} = (x_1, x_2, x_4, x_5)$.

(3)

(ii) Write down the minimal path sets and minimal cut sets.

(4)

(iii) The dual of a system with structure function $\phi(x)$ is defined as the system with structure function

$$\phi_D(x) = 1 - \phi(1-x).$$

Prove that the dual of the system in (i) is

$$\phi_D(x) = (1-x_3)\phi_{1D}(\tilde{x}) + x_3\phi_{2D}(\tilde{x}).$$

(7)

(iv) Sketch the block diagram for the dual system.

(4)

(v) Write down the minimal path sets and minimal cut sets for the dual system.

(2)