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

GRADUATE DIPLOMA IN STATISTICS, 2000

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 25 printed pages followed by 1 blank page (page 26).
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. Why is it sometimes useful to test the null hypothesis that a set of *n* observations can be regarded as a random sample from a Normal distribution, with mean and variance estimated from the observations?

(5)

The probability density function of the Normal distribution with mean μ and variance σ^2 is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right) , \qquad -\infty < x < \infty \ .$$

Given a random sample $x_1, x_2, \ldots, x_n$ from it, find the maximum likelihood estimators of μ and σ^2.

(5)

In order to investigate accusations that their outlets are making excessive profits, a multinational fast-food franchising firm obtains accounts from a random sample of 1000 outlets in a country, and calculates profits as a percentage of turnover for each of them. The results are tabulated as follows:

Profits (%)	*Number of outlets*
1 but less than 2	4
2 but less than 3	13
3 but less than 4	73
4 but less than 5	210
5 but less than 6	357
6 but less than 7	263
7 but less than 8	75
8 but less than 9	5
Total	1000

Test the null hypothesis that the population of outlets' profits is Normal.

(7)

What do you conclude from your analysis?

(3)

Turn over

A2. Statistics of UK exports of goods and services and of gross domestic product, £m at current market prices, 1980-1997, are collected from Table 1.2 of *United Kingdom National Accounts 1998 edn*. They are used to get *y*, the percentage of exports in gross domestic product. A three-year moving average MA is compiled using these percentages. The results are as follows, where "na" indicates "not available":

Year	1980	1981	1982	1983	1984	1985	1986	1987	1988
y	27.22	26.69	26.31	26.45	28.32	28.77	25.69	25.45	23.03
MA	na			27.03	27.85	27.59	26.64	24.72	24.10

Year	1989	1990	1991	1992	1993	1994	1995	1996	1997
y	23.81	24.08	23.22	23.62	25.41	26.44	28.41	29.19	28.52
MA	23.64	23.70	23.64	24.08	25.16	26.75	28.01	28.71	na

Complete the above table by calculating MA for 1981 and 1982. Why is it not possible to obtain MA for 1980 or for 1997?

(2)

Writing t = -8.5, -7.5, …, +8.5 for the years shown above, it may be found that $\Sigma t = 0$, $\Sigma y = 470.63$, $\Sigma t^2 = 484.50$, $\Sigma y^2 = 12374.0495$ and $\Sigma ty = 4.435$. Find the regression of y on t, with r^2 and the conventional estimates of the standard errors of the coefficients.

(6)

Draw a time chart showing both y and MA and draw your regression on it.

(6)

Write a short essay comparing regression and moving average trends, with special reference to this example.

(6)

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

(i) costs of a fixed basket of commodities etc,
(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 Paasche (current-weighted) price index numbers.

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

(8)

(b) 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 number?

(2)

In order to compile this index number, data relating to the expenditure pattern of such households have to be collected by means of interviewing a sample of such 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 sample of such areas is taken and interviewing 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)

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

A4. In order to examine the inter-relationship between male and female employment in Great Britain, quarterly data, seasonally adjusted, of numbers of males and of females in employment (thousands) relating to Spring 1992 to Autumn 1997 inclusive are collected from *Economic Trends Annual Supplement 1998 edn*, Table 3.4. Male numbers are denoted by M, female numbers by F, and a time trend t = 1, 2, …, 23 is taken. It is calculated that

$$\sum M = 323,233 \qquad \sum M^2 = 4,543,780,864 \qquad \sum tM = 3,909,525$$

$$\sum F = 261,961 \qquad \sum F^2 = 2,984,400,128 \qquad \sum tF = 3,170,233$$

$$\sum t = 276 \qquad \sum t^2 = 4,324 \qquad \sum MF = 3,682,407,936 \ .$$

Find the simple correlation coefficients *r(MF), r(tM)* and *r(tF)* and the partial correlation *r(MF.t)*. Test the separate null hypotheses that each of these coefficients is zero in the population from which the data have been drawn as a random sample.

(4)

Find the regression of M on t with estimated standard errors of the coefficients.

(4)

The equivalent regression of F on t is

$$F = \underset{(23.5)}{11073.0} + \underset{(1.715)}{26.384\, t},$$

and the multiple regression of F on M and t is

$$F = \underset{(934.6)}{5783.3} + \underset{(0.06826)}{0.38641\, M} + \underset{(2.342)}{14.651\, t}$$

How do you account for the difference in the constant terms and in the coefficients of t between these two regressions?

(3)

How is the significance of the correlation coefficients (including the partial correlation coefficient) related to the significance of the coefficients in the regressions?

(3)

Find R^2 and $\bar{R}^2$ for the regression of F on M and t.

(2)

What do you learn from this analysis?

(4)

SECTION B - ECONOMETRICS

B1. An economist intends to estimate a regression equation relating demand (y) for a product to its price (x_1) and consumers' income (x_2). This is to be based on 48 quarterly observations extending over 12 years, and to use data that have not been seasonally adjusted. It is believed that the market for this product is seasonal.

(i) A possible approach to account for seasonality is to estimate the model

$$y_t = \alpha + \beta_1 x_{1t} + \beta_2 x_{2t} + \gamma_1 Q_{1t} + \gamma_2 Q_{2t} + \gamma_3 Q_{3t} + \gamma_4 Q_{4t} + \varepsilon_t$$

where ε_t is a random error term and Q_{it} (i = 1, 2, 3, 4) are dummy variables, defined so that Q_{it} takes the value 1 in quarter i of each year and 0 otherwise, i = 1, 2, 3, 4. Explain why there are difficulties in estimating the model by least squares.

(4)

(ii) A model that can be estimated by least squares is

$$y_t = \alpha + \beta_1 x_{1t} + \beta_2 x_{2t} + \gamma_1 Q_{1t} + \gamma_2 Q_{2t} + \gamma_3 Q_{3t} + \varepsilon_t .$$

Interpret the parameters γ_1, γ_2 and γ_3 in this model, and explain how to test the null hypothesis of no seasonal influence on demand.

(8)

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

(8)

B2. (a) The Dickey-Fuller test is often applied as one of the first steps in the analysis of economic time series. What are the null and alternative hypotheses, and why is it believed to be important to test these hypotheses?

Describe the test procedure as it might be applied to real gross domestic product of a national economy.

(12)

(b) It is suspected that at least one cointegration relationship exists among a group of integrated economic time series. Discuss the impact of such a relationship on the vector autoregressive representation.

(8)

B3. Consider the regression model

$$y_i = \beta x_i + \varepsilon_i \qquad (i = 1, ..., n)$$

where

$$E(\varepsilon_i) = 0, \quad E(\varepsilon_i \varepsilon_j) = 0 \quad \text{if } i \neq j$$

and

$$E(\varepsilon_i^2) = \begin{cases} \sigma_1^2 & (i = 1, ..., n_1) \\ \sigma_2^2 & (i = n_1 + 1, ..., n) \end{cases}$$

where n_1 is known and $1 < n_1 < n-1$.

(i) Derive the generalised least squares estimator, $\hat{\beta}$, of β and find the variance of $\hat{\beta}$.

(6)

(ii) Let b denote the ordinary least squares estimator of β. Find the variance of b, and show that b is less efficient than $\hat{\beta}$.

(7)

(iii) Let

$$s^2 = (n-1)^{-1} \sum_{i=1}^{n} (y_i - bx_i)^2 \quad .$$

Find the expected value of s^2 and discuss the testing of hypotheses about β based on standard ordinary least squares methodology.

(7)

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

(a) Probit analysis.

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

(c) Testing for heteroscedasticity in the errors of a regression equation.

(d) The effect of errors in variables on regression estimation.

(e) The Lagrange multiplier test and its econometric applications.

SECTION C - OPERATIONAL RESEARCH

C1. (i) Describe the costs which are commonly assumed in inventory control models. What factors should be considered when determining these costs?

(4)

(ii) The stock-holding cost for a product is £45 per item per annum, and the cost of placing an order for a replenishment is £200. Demand is steady at an annual demand of 2000 items. Shortages must not occur. The purchase cost depends on the number of items Q in the order; the cost per item is £c, where

$$c = \begin{cases} 50 - \dfrac{Q}{100} & \text{for } Q \le 500, \\ 45 & \text{for } 500 \le Q < 1000, \\ 40 & \text{for } Q \ge 1000. \end{cases}$$

Determine the optimal batch order quantity.

(8)

(iii) An electronics company manufactures hi-fi systems, but purchases speakers from a supplier. The cost of each speaker is £50, although this cost is reduced to £45 if at least 2500 speakers are ordered at the same time.

Usage for the speakers is steady, and the monthly usage is 1000 speakers. Shortages must not occur. The cost of placing an order for a replenishment is £500. The company owns a facility which can store up to 200 speakers at a cost of £4 per speaker per month. A warehouse with additional storage space for up to 3000 speakers can be leased at a cost of £2000 per month. Storage costs per speaker in the leased warehouse are also £4 per month.

Should the company lease the warehouse, and what order quantity for speakers should be used?

(8)

C2. (i) Briefly describe five methods which could be used when trying to validate a simulation model.

(8)

(ii) A random variable X has the right-triangular probability density function $f(x)$ with parameters $\alpha < \beta$ as shown below:

$$f(x) = \begin{cases} 0 & x < \alpha \\ \dfrac{2(x-\alpha)}{(\beta-\alpha)^2} & \alpha \le x \le \beta \\ 0 & x > \beta \end{cases}$$

Derive an algorithm for generating observations from this distribution using the method of inversion.

(6)

(iii) Using the uniform distribution on $[\alpha, \beta]$ as the enveloping distribution, derive an algorithm for generating values from $f(x)$ using the method of rejection.

(6)

C3. (i) Consider a queueing system in which customers arrive singly at random at a rate λ. Service times are independent and have an exponential distribution with mean $1/\mu$. Define the *traffic intensity* ρ and write down a formula for L, the expected number of people in the system, in terms of ρ.

(6)

(ii) An engineer is planning the design of a system which will provide an essential service for customers who will arrive singly at random. Service times of customers are independent exponentially distributed random variables, and the mean service time is one of the variables that the engineer can control. The cost of the system increases with the average service rate. However, slow service will result in a queue of customers which is highly undesirable. The engineer knows that the average interval between successive arrivals is 100 seconds, so she has proposed a system design in which the average service time is 95 seconds. She believes this will prevent large queues forming. Do you agree or disagree with this proposal? Support your arguments with explanatory comments.

(4)

(iii) Consider a similar system in which the mean time between successive arrivals is known to be 200 seconds. The system can be designed with a mean service time of 140, 160 or 180 seconds. The cost of running the system for one second depends on the chosen service rate, as shown below.

Mean service time (seconds)	*Cost (£ per second)*
140	100
160	80
180	50

The cost, per second, of congestion in the system is £15L, where L is the mean number of customers in the system. Design an optimal system which will minimise the total running and congestion costs. What assumptions are you making in order to use this model?

(10)

C4. (i) A furniture company manufactures tables, chairs, desks and bookcases. They are made from hardwood and softwood. Each item of furniture takes a certain number of labour-hours to make. The requirements of softwood, hardwood and labour for each item, and the selling prices, are given in the following table.

	Softwood (feet)	*Hardwood (feet)*	*Labour (hours)*	*Price (£)*
Table	5	2	3	31.00
Chair	1	3	2	21.00
Desk	9	4	5	51.50
Bookcase	12	1	10	69.50

The production planning period is 10 days. Up to 10 workers, working an 8-hour day, can be used, at a cost of £5 each per hour. Up to 1500 feet of softwood, at £1 per foot, and up to 1000 feet of hardwood, at £2.50 per foot, are available. Obtain a linear programming formulation of this production planning problem, and find its solution using the simplex method.

(10)

(ii) The Managing Director notices that some items are not being produced at all, and asks by how much the selling prices of these items must be increased before it is worth producing them. Calculate these amounts.

(1)

(iii) If extra hardwood becomes available, would it be worth buying, and how much would it be worth paying for it?

(2)

(iv) Would it be worth putting any workers on overtime at a total cost of £7.50 per hour?

(2)

(v) If it is possible to make sideboards using 15 feet of softwood, 5 feet of hardwood and 12 hours of labour, would it be profitable to make them if the selling price is £115?

(5)

SECTION D - MEDICAL STATISTICS

D1. The survival times from diagnosis (in months) for 18 patients with prostate cancer from one centre are given below.

2*, 14, 23*, 24*, 26, 36, 42, 43*, 51*, 52*, 58*,
59*, 61*, 62*, 65*, 67*, 67*, 69

(* Indicates a right-censored observation.)

(i) Explain what is meant by a left-, right-, and interval censored observation. Specify reasons why left-, right- and interval censoring may occur.

(6)

(ii) Compute the Kaplan-Meier estimate of the survival curve and plot it. Calculate the associated standard error for the Kaplan-Meier survival function estimate using Greenwood's formula at 36 months follow-up.

(10)

(iii) Find an approximate 95% confidence interval for the three-year survival rate for this type of prostate cancer patients.

(2)

(iv) Use your graph to estimate the median survival time for this group of prostate cancer patients.

(1)

(v) The 18 patients were part of a larger clinical trial to compare two treatments, placebo and diethylstilbestrol (DES) for prostate cancer. What analysis could be used to compare the survival times for these two treatments?

(1)

D2. (a) Briefly describe the scope of Phase I, II, III and IV clinical trials.

(4)

(b) A parallel group Phase III multi-centre randomised controlled clinical trial (RCT) protocol is being designed to compare the efficacy of two drugs for the treatment of peptic ulcer. A new drug A is to be compared to the standard treatment drug B. In the RCT the primary outcome is whether or not a patient's ulcer has healed after 40 days of treatment. The anticipated proportions of patients whose ulcers heal after 40 days on the new treatment (drug A) and standard treatment (drug 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 to receive the standard drug B. The null hypothesis is that $p_a = p_b = (P_1 + P_2)/2$ 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 type I error (α) and type II error (β), using a two tailed test.

(8)

(ii) The ulcer healing rate of patients after 40 days of treatment on the standard drug B is approximately 60%. The new drug A would be considered effective if it increased the ulcer healing rate to 80%. Evaluate n for α (two sided) = 0.05, $\beta = 0.20$.

(2)

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

(6)

D3. (a) Discuss the advantages and disadvantages of a cohort study compared to a case-control study for the evaluation of the association between smoking and lung cancer.

(10)

(b) Doll and Hill (1950) carried out a case-control study into the aetiology of lung cancer. There were 709 lung cancer patients and 709 controls drawn from the same hospitals who were compared for their smoking history.

Numbers of smokers and non-smokers among lung cancer patients and age and sex matched controls with diseases other than cancer.

	Non-smokers	**Smokers**	**Total**
Males:			
Lung cancer patients	2	647	649
Control patients	27	622	649
Females:			
Lung cancer patients	19	41	60
Control patients	32	28	60

(Source: Doll and Hill, 1950, British Medical Journal)

(i) Calculate the Mantel-Haenszel estimate of the odds ratio for lung cancer due to smoking, allowing for sex. Calculate a 95% confidence interval for this odds ratio.

(6)

(ii) Perform a test of the null hypothesis that smoking is unrelated to lung cancer.

(3)

(iii) Comment on your findings in parts (i) and (ii).

(1)

D4. (a) Define the sensitivity and specificity of a diagnostic test. Derive expressions for positive predictive value and negative predictive value in terms of sensitivity, specificity and prevalence.

(4)

(b) The table below shows the results of an HIV antibody assay among patients with AIDS and healthy blood donors (without AIDS). The results are expressed as the ratio of the mean absorbance of a pair of test samples divided by the mean absorbance of eight negative control wells.

Ratio	*Healthy blood donors*	*Patients with AIDS*	*Total*
< 2.0	202	0	202
2.0 - 2.99	73	2	75
3.0 - 3.99	15	7	22
4.0 - 4.99	3	7	10
5.0 - 5.99	2	15	17
6.0 - 11.99	2	36	38
12.0 +	0	21	21
Total	297	88	385

(Source: Weiss et al, 1985, Journal of the American Medical Association)

(i) Six potential cut-off values for a diagnostic test are 2.0, 3.0, 4.0, 5.0, 6.0 and 12.0. Determine the corresponding test sensitivities and specificities.

(8)

(ii) Determine the positive and negative predictive values of the tests in populations in which the prevalence of AIDS is 1%.

(4)

(iii) Sketch the ROC curve using the six cut-off values from part (i).

(4)

SECTION E - BIOMETRY

E1. Explain the importance of randomisation in the design and analysis of experiments.

(4)

Potato cyst-nematode (PCN) is a serious pest of potatoes. It can be controlled by chemical nematicides, by deep cultivation, or by planting varieties resistant to PCN. In a field experiment to study the effects of these methods of control, all eight combinations of three factors, each at two levels, were laid out in each of three randomised blocks on nematode-infested land.

The factors were as follows:

Variety: Resistant (R) v Susceptible (S)
Cultivation: Deep (D) v Normal (-)
Nematicide: None (-) v Nematicide (N)

The mean yields of potato tubers (tonnes/hectare) for each treatment combination were as follows:

Variety	*Cultivation*	*Nematicide*	*Mean Yield*
R	D	-	29.2
R	D	N	42.2
R	-	-	25.8
R	-	N	30.4
S	D	-	4.3
S	D	N	35.4
S	-	-	5.6
S	-	N	25.0

Complete the analysis of variance for all main effects and interactions, given that the residual SS was 221.6. Interpret the result.

(12)

Compute the standard error of

(i) the difference between the means of the two varieties,

(ii) the mean effect of nematicide for a particular variety.

(4)

E2. Explain the meaning of the terms *quantal response* and *logistic regression*.

(6)

Several pyrethroid insecticides were compared in a study of the effects of insecticides on houseflies. Groups of houseflies were exposed to different concentrations of insecticide, and the numbers lying motionless after 30 minutes were recorded. The results for two particular insecticides, A and B, were as follows:

Concentration	**Number tested**	**Number motionless**
Insecticide A		
1.28	30	0
3.20	53	4
8.00	50	15
20.00	49	46
50.00	48	48
Insecticide B		
3.20	25	0
8.00	52	11
20.00	49	27
50.00	43	41

Plot the data on transformed scales, with log concentration as the *x*-axis and a suitable transformation of the proportion motionless as the *y*-axis. Comment on your graph.

(6)

The results of a logistic regression of the proportion motionless against log concentration (logs to base 10) were as follows:

	Insecticide A	Insecticide B
Intercept	–6.886	–6.322
Slope	7.065	5.209
Residual deviance	4.507	3.268

Assuming common slope:

Intercept	–5.930	–7.368
Slope	6.075	6.075
Combined Residual Deviance	10.197	

When both sets were combined the residual deviance was 29.816.

Question E2 continued on next page

Estimate the relative potency of insecticide A compared with insecticide B.

(4)

Explain how you would test the hypotheses

(i) that the slope is the same for each insecticide,

(ii) that the insecticides differ in potency.

(4)

E3. Explain the importance of balance in the design of experiments.

(5)

An experiment to compare six varieties, A-F, of a particular crop was laid out in four randomised blocks according to the following plan:

Block 1	C	F	A	E	B	D
Block 2	B	E	D	A	F	C*
Block 3	E	C	A	B	D	F*
Block 4	F	B	E	D	C*	A*

Due to floods, the plots marked with asterisks could not be harvested.

You are presented with the results and asked to provide the best analysis possible.

(i) Describe in detail how you would estimate the block and variety means, using a method for estimating missing plots.

(5)

(ii) Outline the analysis of variance table, showing how you would test for the significance of differences between varieties.

(5)

(iii) Give formulae for the standard errors of the difference between varieties A and B, and between varieties A and C.

(5)

E4. Describe the Newton-Raphson method for fitting nonlinear regression models by the method of least squares.

(6)

The response of the yield of cereals (Y) to increasing applications of nitrogen fertiliser (N) is often modelled by the negative exponential curve

$$Y = \alpha + \beta \exp(-\kappa N)$$

where α, β and κ are unknown parameters to be estimated.

In a continuous experiment in which Spring Barley was grown on the same plots in successive years, the mean yields (in tonnes/hectare) over plots with four levels of application of nitrogen fertiliser were as follows:

Year	*Levels of nitrogen*			
	0	1	2	3
1	1.54	2.93	3.66	3.78
2	2.17	2.54	2.68	2.88
3	2.33	3.92	4.61	4.69

(i) Assuming that each mean yield is independently distributed with equal variance, describe how you would use the Newton-Raphson method to fit three exponential curves to the responses, one for each of the years.

(4)

(ii) How would you obtain initial estimates for the parameters?

(4)

(iii) The fitted parameters were as follows:

Year 1:	$\hat{\alpha} = 4.00$	$\hat{\beta} = -2.47$	$\hat{\kappa} = 0.877$
Year 2:	$\hat{\alpha} = 3.12$	$\hat{\beta} = -0.95$	$\hat{\kappa} = 0.432$
Year 3:	$\hat{\alpha} = 4.85$	$\hat{\beta} = -2.52$	$\hat{\kappa} = 1.044$

How would you construct a simpler model which would make it easier to explain the differences between the response curves in each year, and how would you fit such a model?

(6)

SECTION F - STATISTICS FOR INDUSTRY AND QUALITY IMPROVEMENT

F1. A quality improvement team in a car factory wants to improve productivity. Team members decide to look at numbers of stoppages recorded monthly over the last two years. The data are shown below:

Month	**Number of Stoppages**	
	1997	*1998*
January	12	7
February	9	2
March	3	8
April	7	11
May	7	7
June	6	7
July	8	10
August	7	3
September	10	6
October	9	8
November	13	4
December	6	11

Throughout this question you should assume that months have equal length.

(i) The team starts by setting up a one at a time control chart of the data.

(a) Check whether the Poisson model is plausible for these data. (2)

(b) Is there any evidence of a difference in mean stoppages between the two years? (2)

(c) Assuming that the process is in statistical control during 1997 and 1998, suggest suitable control limits for a one at a time control chart of the data, and demonstrate its use for the following new monthly data:

Jan	Feb	Mar	April	May	June	July	Aug	Sept	Oct
13	14	18	11	8	12	8	23	19	9

Are there any action signals? (4)

(ii) The team decides to use a 3-month moving average control chart as well, because data are also discussed at quarterly Board meetings.

(a) Suggest approximate control limits for the moving average chart, and demonstrate its use for the new monthly data given in part (i)(c). (4)

(b) Give one advantage and one disadvantage of the moving average chart compared with the one at a time chart. (1)

(iii) The team decides to use a Cusum chart for future monthly data.

(a) Set up a Cusum chart about a level of 8 and plot the new monthly data given in part (i)(c). (4)

(b) Demonstrate the use of a *V*-mask after the August datum. (2)

(iv) What other information should be considered as well as the number of stoppages? (1)

F2. In an experiment on the final stage of a polymerisation process, the temperature is kept constant throughout a batch and samples are taken off after 35, 40 and 45 minutes. The melt flow index (MFI) of each sample is measured, and the results are given below.

		Time (mins)		
Batch	**Temp (°C)**	**35**	**40**	**45**
A	175	3.19	3.45	3.57
B	185	4.21	4.22	4.22
C	185	3.58	3.76	3.84
D	175	2.92	3.34	3.42
E	185	3.92	3.93	3.96
F	175	3.45	3.68	3.72

(i) Consider the MFI after 45 minutes. Construct a 95% confidence interval for the mean difference in operating the process at 175°C and 185°C. State any assumptions you make.

(6)

(ii) The model

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_6 x_{6i} + \beta_7 x_{1i}^2 + E_i$$

is fitted where x_1 is time, coded as –1, 0 and 1 for 35, 40 and 45 minutes respectively, and $x_2, \ldots, x_6$ are indicator variables that identify batch effects relative to batch F, that is

	x_2	x_3	x_4	x_5	x_6
A	1	0	0	0	0
B	0	1	0	0	0
C	0	0	1	0	0
D	0	0	0	1	0
E	0	0	0	0	1
F	0	0	0	0	0

The E_i are independent random errors. An excerpt from a regression analysis follows.

coefficient	**estimate**	**standard error**
β_0	3.6589	0.06829
β_1	0.12167	0.02957
β_2	–0.21333	0.08364
β_3	0.60000	0.08364
β_4	0.11000	0.08364
β_5	–0.39000	0.08364
β_6	0.32000	0.08364
β_7	–0.06333	0.05122

Question F2 continued on next page

(a) What is meant by the statement that x_1 is orthogonal to x_1^2 ?

(1)

(b) What is the correlation between x_2 and x_3?

(2)

(c) Suppose temperature is coded –1 and +1 for 175ºC and 185ºC respectively and denoted by x_7. What is the correlation between the linear combination

$$0.5 + 0.5x_7 - x_3 - x_4$$

and x_6? Hence explain why x_7 cannot be added to the regression model. Make an estimate of the temperature effect from the estimated coefficients of the indicator variables, and construct a 95% confidence interval for this effect. State clearly all the assumptions you are making.

(7)

(d) Estimate the between batch variance, and hence the between batch standard deviation.

(4)

F3. From each of 6 ingots, 3 discs were prepared. For each set of 3 discs, one disc was subjected to heat treatment *A*, one to heat treatment *B* and the third to heat treatment *C*. The allocations of discs to treatments were randomised. Yield points in Newtons per mm^2 for the treated specimens are given below:

		Ingot number					
		1	*2*	*3*	*4*	*5*	*6*
Treatment	*A*	347	292	301	273	310	349
	B	344	308	345	317	345	351
	C	356	332	355	320	339	326

(i) Write down a suitable model for yield points which would be appropriate if the ingots were considered to be random effects.

(2)

(ii) The between treatments, between ingots and total (corrected) sums of squares are 2428.0, 4765.3, 9776.0 respectively. Write down the ANOVA table including the expected values of the mean squares.

(3)

(iii) Using a 5% significance level, test the hypothesis that there is no difference between treatments.

(3)

(iv) Estimate the between ingots variance, and hence the between ingots standard deviation.

(2)

(v) Calculate the mean yield point for each treatment and the least significant difference, LSD (5%), for the differences between means. Comment on your results.

(3)

(vi) What is the difference between a randomised block design and a completely randomised design with two factors? Suppose ingots 1 up to 6 contained different additives. Would you advise analysing the data as though they were from a completely randomised design with two factors, heat treatment and additive? Give a reason for your answer.

(3)

(vii) Variability between ingots is considered a noise factor, and it is as important to minimise the variability of yield points, for a particular treatment, as it is to maximise the average yield point. Can you make any inferences about variability from these data? How might the experiment have been extended to provide more information about variability?

(4)

F4. (a) A factory has 10 identical machines which were all overhauled two years ago. Seven machines have since failed. The times in months until first failure were:

2 8 10 11 13 18 18 .

Use a graphical method to estimate the parameters, α and β, in the Weibull distribution with cumulative distribution function

$$F(t) = 1 - \exp(-(t/\beta)^{\alpha}) \quad .$$

You may assume that

$$F\left(t_{(i)}\right) \approx \frac{i - 0.4}{n + 0.2}$$

where $t_{(i)}$ is the ith order statistic. Hence estimate the probability that all machines have failed within 3 years of the overhaul.

(10)

(b) In a twin-engined plane, for either engine, the unconditional probability that it fails at some time during a flight is q. However, engine failures are not independent:

P(right engine fails at some time during a flight | left engine fails at some time during the flight) = a

and similarly for the left engine failing given that the right fails at some time. The probability that one engine fails when the other does not is given by

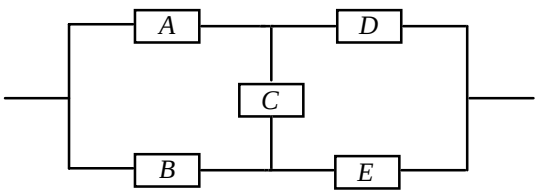
P(right engine fails | left engine does not fail) = b

and similarly for the left engine failing given that the right does not.

Determine the relation between a, b and q. Find the probabilities of 0, 1 and 2 engines failing in terms of a and q.

(5)

(c) What is the overall reliability of the system



if components have reliabilities R_A, R_B, R_C, R_D and R_E, and failures are independent?

(5)

BLANK PAGE