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

GRADUATE DIPLOMA IN STATISTICS, 1996

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.

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.*

SECTION A – STATISTICS FOR ECONOMICS

A1. (i) One moving average trend which may be fitted to a series of observations $\ldots x_{t-2}, x_{t-1}, x_t, x_{t+1}, x_{t+2} \ldots$ is $y_t = x_{t-2}/8 + x_{t-1}/4 + x_t/4 + x_{t+1}/4 + x_{t+2}/8$. If the observations have been generated by a stationary process in which $\rho(x_t, x_{t+i}) = 0$ for all $i > 0$, find the autocorrelation coefficients $\rho(y_t, y_{t+i})$ for $i = 1, 2, \ldots$.

(ii) Suppose that, instead of being a stationary process, the observations are of a quarterly economic time series. What advantage does this moving average trend have in this instance?

Consider the following data:

Manufacturers' sales of floorcoverings
UK, millions of square metres

1987	1st quarter	37.0	1990	1st quarter	39.8
	2nd quarter	39.5		2nd quarter	37.3
	3rd quarter	39.7		3rd quarter	37.5
	4th quarter	45.4		4th quarter	38.2
1988	1st quarter	40.3	1991	1st quarter	32.9
	2nd quarter	41.7		2nd quarter	34.0
	3rd quarter	41.5		3rd quarter	35.3
	4th quarter	45.3		4th quarter	37.9
1989	1st quarter	40.8	1992	1st quarter	35.4
	2nd quarter	38.5		2nd quarter	33.0
	3rd quarter	37.6		3rd quarter	34.6
	4th quarter	41.3		4th quarter	36.9

(Source: Monthly Digest of Statistics, June 1995, Table 11.6)

Calculations were performed on them using the Minitab Statistics package as shown on the next page of this examination paper.

(iii) Making such use of the computer output as you wish, correct the data for 1991 and 1992 for seasonal effects using the ratio-to-a-moving-geometric-average method.

(iv) Why would the differences-from-a-moving-arithmetic-average method be less satisfactory?

(v) Interpret the estimated seasonal pattern in the above data in percentage terms.

A2. Statistics of the exports of goods, £m at 1990 prices for each year between 1972 and 1993 inclusive, are collected from Table 1.3 of United Kingdom National Accounts, 1994 edition, and are denoted by X. Their natural logarithms, correct to two decimal places, are denoted by y so that $y = \log_e X$. The variable t is defined as taking the values -10.5, -9.5 ... 10.5 over the period, so there are 22 pairs of observations (t, y). It is found that $\Sigma t = 0,\ \Sigma t^2 = 885.5,\ \Sigma y = 246.39,\ \Sigma y^2 = 2760.8571$ and $\Sigma ty = 34.895$.

Estimate α and β in the model $y = \alpha + \beta t + u,$ where $u \sim NID(0,\ \sigma^2)$, by ordinary least squares. Obtain the coefficient of determination (r^2) and estimate σ^2 and the standard errors of the coefficients.

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.

Use your estimated model to predict y for 1994.

Two different standard errors are associated with such predictions, and hence two different 95 per cent confidence intervals. Explain what these intervals relate to, and calculate them for your prediction.

What is the model giving X as a function of t equivalent to your estimated model? What value of X for 1994 is predicted by your estimated model? What are the two corresponding 95 per cent confidence intervals, in numerical terms?

Turn over

A3. (a) Samples of shares listed in the Financial Times of 18 July 1995 were taken, and their gross yields (%) noted as follows:

Food retailers: 1.4 2.9 3.5 0.0 3.2 4.4
sum = 15.4 sum of squares = 52.22

General retailers: 3.9 0.0 4.2 3.0 2.9 5.3 5.7 0.0
sum = 25.0 sum of squares = 110.84

Leisure and hotels: 4.3 6.3 0.0 2.3 0.0 0.8 5.2 1.5 1.2 4.9
sum = 26.5 sum of squares = 118.85

Use an analysis of variance to test the null hypothesis that means in the three populations from which the samples were drawn were the same.

On what assumptions is your test based?

Are there any reasons to doubt the validity of any of the assumptions?

(b) Samples of unit trusts and investment trusts were taken and the net returns (%) on them over a twelve month period were calculated:

Unit trusts: 7.7 3.5 2.6 5.0 5.6 4.6 0.2 2.2 3.2 4.2

sum = 38.8 sum of squares = 188.58

Investment trusts: 4.1 5.3 3.3 11.0 7.3 7.0 5.2 4.9 3.7 6.3 2.4 3.8 4.5 2.6 9.8

sum = 81.2 sum of squares = 526.76

Assuming that the standard deviations of net returns of unit trusts and of investment trusts in the populations from which the samples were drawn were equal, estimate the difference of the mean net returns in the population from which the samples were drawn, together with a 95% confidence interval for your estimate.

Does your confidence interval indicate anything about the result of an analysis of variance which might be carried out on the same data? [Do not carry out such an analysis of variance.]

The assumption of equal standard deviations of net returns could be challenged on the grounds that changes in investment trusts' discounts would make one expect that their standard deviation would be rather larger than that of unit trusts. How might you modify your analysis to take this possibility into account?

A4. (a) Explain, using formulas, *Laspeyres* and *Paasche price indices*. At times of rising prices, which is usually expected to show the larger increase? Why?

Why are Laspeyres indices more commonly used in practice than Paasche indices?

(b) The official Index of Retail Prices in a particular developing country with a largely agricultural economy is published monthly. However, it has been subjected to widespread criticisms with the result that it does not command public confidence. You have been invited to visit the country to investigate whether the Index is properly constructed and to make recommendations about how it might be improved. How would you go about your work, and what problems would you expect to encounter?

SECTION B – ECONOMETRICS

B1. Consider the following structural model of two equations:

$$\left.\begin{aligned} y_{1t} + \gamma_{21} y_{2t} + \beta_{11} x_{1t} &= \varepsilon_{1t} \quad (1) \\ y_{1t} + \gamma_{22} y_{2t} &= \varepsilon_{2t} \quad (2) \end{aligned}\right] \; t = 1, \ldots, n$$

where the error terms $(\varepsilon_{1t}, \varepsilon_{2t})$ are not autocorrelated but may be contemporaneously correlated, (y_{1t}, y_{2t}) are endogenous and x_{1t} is exogenous. Also assume that $\gamma_{21} \neq \gamma_{22}$.

(a) Derive the reduced form equations.

(b) Provide necessary and sufficient conditions for (2) to be identified.

(c) What role is played by the assumption $\gamma_{21} \neq \gamma_{22}$?

(d) Assuming that (2) is identified, describe how γ_{22} can be estimated by two-stage least squares.

(e) Equation (1) is not identified. Follow through the steps of two-stage least squares, and show what will occur if you attempt to estimate the parameters of (1) by that method.

B2. (a) What is meant by heteroscedasticity in the errors of a regression equation? Provide an example of an economic model in which this problem might occur.

(b) Discuss a test for heteroscedasticity, including a discussion of the alternative hypothesis for which the test is appropriate.

(c) Describe a possible procedure for estimating a regression model in the presence of heteroscedasticity.

(d) Suppose that heteroscedasticity is present but is ignored, so that the regression is estimated by ordinary least squares. What would be the consequences?

B3. (a) What is meant in saying that the generating model of a time series has a unit autoregressive root? Outline the Dickey-Fuller test for unit roots, and discuss circumstances in which the test might be unreliable.

(b) The series Y_t and X_t denote respectively short- and long-term interest rates. Each series can be assumed to be generated by a process with a single unit autoregressive root. An economist used ordinary least squares to estimate the regression

$$Y_t = \alpha + \beta X_t + \varepsilon_t$$

Denote by $\hat{\varepsilon}_t$ the residuals from the estimated regression. A Dickey-Fuller test was applied to the series $\hat{\varepsilon}_t$. Fully discuss what can be learned from the test outcome. In particular, outline the implications of rejection of the null hypothesis.

B4. An economist is interested in whether, among unmarried applicants for mortgages, there is any difference between the proportions of males and of females who are successful. Data are available on five hundred mortgage applications from single people. Information includes whether the application was successful, the sex of the applicant, and relevant economic and other variables such as the applicant's income and the age of the property. In these cases, there were no substantial differences in either the size of loan requested or the ratio of loan requested to the value of the property. Discuss, as fully as possible, how you would proceed in analysing these data. In particular, your discussion should include

(i) the specification of an appropriate model with associated assumptions,

(ii) an outline of the methodology for estimating this model, and

(iii) a discussion of methods of testing the hypothesis of interest.

SECTION C - OPERATIONAL RESEARCH

C1. (a) Write down the differential-difference equations for a single-server queuing system where the arrival rate λ is constant but the mean service rate μ_n depends on *n*, the number of customers in the system. Hence derive the steady-state expression for P_n, the probability that there are *n* customers in the system, in terms of λ, μ_n, and P_0.

What assumption are you making about the queuing system in order to derive these differential-difference equations?

(b) Consider a queuing system with customers arriving at random at a rate of 15 per hour. When there is only one customer in the system, he or she is served by a single server with a mean service time of 3 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 2 minutes. In both cases the service time is exponential. What is the expected number of customers in the system when it is in steady state?

Hint: you can use the identity

$$\sum_{n=0}^{\infty} nx^n = \frac{x}{(1-x)^2} \qquad \text{for } |x| < 1.$$

C2. (a) A project consists of the following activities, whose prerequisites and durations are given in the table below. Draw a network, suitable for analysis by the critical path method, to represent the project.

For each event, write the earliest and latest event times and deduce the critical path. What is the total float on each task?

Activity	**Prerequisites**	**Duration (days)**
A	-	3
B	-	6
C	-	4
D	A	5
E	B	9
F	B, C	8
G	E	9
H	D, E	4
I	D, E	5
J	F, G	2
K	I, J	3
L	F, G	9
M	H, K	2
N	F	8
O	L, M	2

(Question continued on next page)

(b) It is required to reduce the overall project duration to 30 days. The durations of activities B, G, K and L can be reduced by investment in extra resources. The reduced durations and the costs of making a reduction are shown in the following table. The entire reduction must be made: intermediate reductions of these activities are not possible. What activities do you recommend should be assigned extra resources?

Activity	Minimum duration (days)	Cost (£)
B	3	600
G	7	500
K	1	200
L	6	300

C3. (a) The stock holding cost for a product is £12 per item per annum, and the cost of placing an order for a replenishment is £60. Demand is steady and the annual demand is 1000 items. Shortages must not occur. The standard purchase cost is £50 per item. However a discount of 2.5% is given if 500 or more items are purchased. Alternatively, a discount of 5% is given if 1000 or more items are ordered. Determine an optimal ordering policy.

(b) A boys' football club has reached the final of the Under 12's Cup and has decided to order some special T-shirts for their supporters to wear at the match. The T-shirts have to be ordered in sets of five. The probabilities of the likely demand for these T-shirts have been estimated and are given by the following table.

Demand	Probability
0	0.00
5	0.05
0	0.10
15	0.15
20	0.15
25	0.20
30	0.20
35	0.10
40	0.05
$\geq$45	0.00

The club has to pay the supplier £3.50 for each T-shirt, and then sells them to the supporters for £6.00 each. After the match has taken place, the club could sell off any remaining T-shirts at the reduced price of £2.50 each. However, if the club does not order sufficient T-shirts, the disappointed supporters will refuse to buy tickets for the prize draw, and the club manager estimates that this will result in a potential loss of £2.00 for each supporter who was unable to obtain a T-shirt.

Assuming that all sales will be in multiples of five, how many T-shirts should the club order, to maximise the expected profit?

C4. (a) A bicycle manufacturer produces men's and women's models. Demand for the next two months is shown in the following table.

Model	**Month 1**	**Month 2**
Men's	150	200
Women's	125	150

For each bicycle, the production cost, the time required by the labour force for manufacture and the time required by the labour force for assembly are shown in the following table: the current inventory levels (at the start of month 1) are also listed:

Model	**Production cost (£)**	**Time for manufacture (hours)**	**Time for assembly (hours)**	**Current inventory**
Men's	60	12	4	15
Women's	45	10	3	25

Last month, the company used a total of 4000 hours of labour. The company's labour relations policy will not allow the combined total hours of labour (manufacture plus assembly) to increase or decrease by more than 500 hours from month to month.

There are end-of-month inventory holding costs. For each bicycle in stock at the end of a month, the holding cost is 3% of its production cost. The company requires at least 25 bicycles of each model to be in stock at the end of the second month.

Write down a linear programming formulation (*but do not attempt to solve it*) for the problem of planning production so that demand is satisfied at minimum total cost. You may ignore any requirements for variables to be integer-valued.

(b) The factory management is currently involved in negotiations with the trade unions about the possibility of changing the labour relations policy, so that the maximum allowable monthly change in total labour hours might no longer be 500. How could you use the above linear programming model to investigate the effects of such a change on the total cost?

SECTION D – MEDICAL STATISTICS

D1. What are the advantages and disadvantages of the *cross-over design* relative to the *parallel group design* in a clinical trial to compare two treatments?

In a small trial to assess a new anti-depressant drug, each of sixteen patients received a month's treatment with the drug and a month's treatment with a placebo, the order of receiving the treatments being selected at random. Depression scores were recorded at the end of each treatment period. The scores, which fall in a range from 0 (no depression) to 30, are tabulated below; they may be assumed to be normally distributed.

Group 1

Patient	*Drug (first)*	*Placebo (second)*
1	11	19
2	11	15
3	22	28
4	19	21
5	7	13
6	7	9
7	6	12
8	8	9

Group 2

Patient	*Placebo (first)*	*Drug (second)*
9	20	14
10	16	16
11	22	16
12	6	3
13	16	14
14	11	8
15	24	23
16	12	7

Is there evidence of (a) a period effect, (b) a treatment x period interaction, (c) a treatment effect?

Discuss problems in the interpretation of these results.

D2. In 1983 legislation was introduced requiring all car drivers and front seat passengers to wear seat belts. Using the Table below, discuss whether the legislation was an effective safety measure. Describe any limitations you see in the data, and any other information you would like in order to draw firmer conclusions.

	Killed or seriously injured		Killed	
Year	1982	1984	1982	1984
Car drivers	19,460	16,421	1,472	1,228
Car front seat passengers	9,458	7,047	658	539
Car rear seat passengers	4,706	5,062	297	372
Pedestrians	18,963	19,168	1,869	1,821
Cyclists	5,967	6,506	294	337

Road accidents: number killed or seriously injured, and numbers killed, in Great Britain in 1982 and 1984.

(*Source of data:* *Harvey A.C & Durbin J. (1986)*
The Effects of Seat Belt Legislation on British Road Casualties: a Case Study in Structural Time Series Modelling
JRSS (A); ***149****: 187-227*)

D3. Canner (Statistics in Medicine, 1987) reports the results of the six known randomised, placebo-controlled clinical trials of aspirin to prevent death in people who have recently suffered a heart attack. They are given in the following table.

Study	**Aspirin**		**Placebo**	
	No. of Patients	*No. of Deaths*	*No. of Patients*	*No. of Deaths*
1	615	49	624	67
2	758	44	771	64
3	317	27	309	32
4	832	102	850	126
5	810	85	406	52
6	2267	246	2257	219

Use the Mantel-Haenszel procedure to find an estimate and confidence interval for the odds of death for the placebo group relative to the aspirin group. What factors need to be borne in mind when generalising the results?

Turn over

D4. Say what is meant by *informative* and *non-informative censoring*, and give examples of each.

Twelve of twenty patients diagnosed as having cancer of the pancreas have so far died. The twenty survival times (in months) are:

1, 2, 2, 3, 4, 5*, 6, 8*, 9, 12*, 14, 19*, 22, 30*, 35, 48*, 56, 77*, 90, 124*

where the asterisk denotes that the patient is still alive at the time at which the analysis is required.

For survival times up to and including 9 months, calculate the Kaplan-Meier estimate of the survival function and the associated standard error using Greenwood's formula.

SECTION E - BIOMETRY

E1. An experiment was carried out to study the effect of five substances containing nitrogen on the yield of sugar beet. Six treatments were used, namely

(1) O = No application of nitrogen
(2) A = $(NH_4)SO_4$ a salt of ammonia (inorganic)
(3) B = NH_4NO_3 a salt of ammonia (inorganic)
(4) C = $CO(NH_4)_2$ an organic compound
(5) D = $Ca(NO_3)_2$ a metallic salt (inorganic)
(6) E = $NaNO_3$ a metallic salt (inorganic)

Treatments (2)-(6) were applied so as to give 100 units of nitrogen per acre. Yields were recorded in tons per acre.

The experiment was a randomised complete block design involving six blocks. However, three of the plots became waterlogged during the experiment and their yields were discarded.

Below is given an extract from a computer analysis in which the missing values are estimated and the degrees of freedom for the residual sum of squares are appropriately adjusted.

Assuming an orthogonal analysis, write down contrast coefficients appropriate for a set of meaningful orthogonal contrasts and partition the treatment sum of squares into appropriate components. Complete the analysis of variance and comment on the results obtained.

(Question continued on next page)

***** Analysis of variance *****

Variate : yield

Source of variation	d.f.(m.v)	s.s	m.s.	v.r.	F pr
block	5	36.896	7.379	3.63	0.015
treat	5	143.398	28.680	14.12	<.001
Residual	22(3)	44.695	2.032		
Total	32(3)	201.795			

***** Tables of means *****

Variate: yield

Grand mean 29.88

block	1	2	3	4	5	6
	30.63	28.88	28.28	29.92	31.25	30.29
treat	1	2	3	4	5	6
	25.52	31.11	30.35	31.52	30.38	30.37

Part of a nonorthogonal data analysis excluding the waterlogged plots appears below. Explain how the adjusted mean difference for the contrast found to be the most important in the analysis above may be estimated and how its standard error may be calculated.

*** Estimates of regression coefficients ***

	estimate	s.e.	t
Constant	26.280	0.825	31.86
block 2	-1.750	0.823	-2.13
block 3	-2.350	0.823	-2.86
block 4	-0.717	0.823	-0.87
block 5	0.614	0.871	0.71
block 6	-0.345	0.939	-0.37
treat 2	5.592	0.923	6.06
treat 3	4.827	0.873	5.53
treat 4	5.994	0.873	6.87
treat 5	4.860	0.901	5.39
treat 6	4.844	0.873	5.55

covariance matrix

1	0.6805					
2	-0.3386	0.6772				
3	-0.3386	0.3386	0.6772			
4	-0.3386	0.3386	0.3386	0.6772		
5	-0.3549	0.3386	0.3386	0.3386	0.7588	
6	-0.2692	0.3386	0.3386	0.3386	0.3304	0.8812
7	-0.4234	0.0000	0.0000	0.0000	0.0857	-0.1101
8	-0.4071	0.0000	0.0000	0.0000	0.0041	-0.1020
9	-0.4071	0.0000	0.0000	0.0000	0.0041	-0.1020
10	-0.4063	0.0000	0.0000	0.0000	0.0000	0.0000
11	-0.4071	0.0000	0.0000	0.0000	0.0041	-0.1020

	1	2	3	4	5	6
7	0.8518					
8	0.4275	0.7620				
9	0.4275	0.4234	0.7620			
10	0.4063	0.4063	0.4063	0.8126		
11	0.4275	0.4234	0.4234	0.4063	0.7620	

	7	8	9	10	11

Turn over

E2. (a) A multiple linear regression model of the form $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ has been fitted to a set of data where $\boldsymbol{y}$ is an (n x 1) vector of observations, $\boldsymbol{\beta}$ is a (p x 1) vector of regression coefficients, $\boldsymbol{X}$ is an (n x p) design matrix of the regressor variables and $\boldsymbol{\varepsilon}$ is an (n x 1) vector of uncorrelated errors each having a mean of zero and a variance of σ^2.

(i) Write down the least squares estimator of $\boldsymbol{\beta}$, show that it is unbiased and derive its variance.

(ii) Write down an estimator of the mean value of $\boldsymbol{y}$ at $\boldsymbol{x} = \boldsymbol{x}_a$ and derive its variance.

(b) Multiple regression was used to model the relationship between the activity of cholinestrase in eggs (y) and

(i) the age in days of the egg at the time of treatment (x_1);

(ii) the age in days of the egg at the time of observation (x_2);

(iii) the dose level of the treatment in μg/egg (x_3).

The dose level of the treatment was found to be non-significant and the model was refitted with only x_1 and x_2. From this analysis the fitted model was

$$y = 8.109 - 7.363x_1 + 15.315x_2.$$

The analysis also gave an unbiased estimator, s^2, of the underlying variance as 220.73 on 47 degrees of freedom and

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{bmatrix} 1.3234 & -0.0287 & -0.1449 \\ -0.0287 & 0.0075 & -0.0014 \\ -0.1449 & -0.0014 & 0.0192 \end{bmatrix}$$

Calculate a 95% confidence interval for the predicted (mean) activity in eggs which were 5 days old at the time of treatment and 8 days old at the time of observation when treatment was given at a dose level of 0.6 μg/egg.

Another statistician carried out a similar analysis, except that she used the full model with all three explanatory variables. Her analysis gave a value of s^2 of 223.19 and she calculated the predicted mean value to be 95.6 with a standard error of 3.41. Comment critically on how your approach and results compare with hers.

E3. Explain what is meant by a *non-linear model*. Give a biometrical example of a nonlinear model that can be transformed to a linear model. How would you decide whether such a transformation was appropriate?

A set of data on the nutrient uptake (y) over time (t) by a cell in a constant medium can be described by the model

$$y_i = \alpha(1 - e^{-\beta t_i}) + \varepsilon_i, \ (i = 1, 2, \ldots, n).$$

The experimenter wishes to use the method of least squares to estimate the parameters. Write down the expression to be minimised, and obtain the normal equations. Use these to illustrate why the estimation of parameters in nonlinear modelling has to be done iteratively and show briefly how you could do it in this situation.

The following data have been collected and a model of the form $y = \alpha(1 - e^{-\beta t})$ is to be fitted.

t	1.5	2.5	4.0	5.0	7.0	9.5	13.0	16.5	22.5	29.0	33.5
y	0.06	0.25	0.47	0.35	0.57	0.62	0.67	0.84	0.90	0.84	0.87

Plot the data. Using your plot, derive initial estimates of α and β.

E4. Two new formulations (*A* and *B*) of a drug have been developed. A bioassay has been designed to estimate the potency of each of the new formulations relative to the standard formulation. Explain what is meant by *relative potency*.

The bioassay is a parallel line assay in which the errors in the response variable are assumed to be normally distributed. Several different doses of formulation *A*, formulation *B* and the standard formulation have been tested. Explain how an estimate of relative potency can be obtained.

The assay measured blood sugar levels in 77 individual mice. Groups of mice were treated at 4 doses of the standard formulation, 3 doses of formulation *A* and 4 doses of formulation *B*. Two general linear models were fitted to the data giving the output shown on the next page. DRUG is a factor relating to the formulation of the drug and LDOSE is the natural logarithm of the dose given to each mouse. Are these results consistent with a parallel line assay? If so, estimate the potency of formulation *A* and formulation *B* relative to the standard preparation.

(Question continued on next page)

General Linear Models Procedure

Dependent Variable: RESPONSE

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	65.490146	13.098029	12.58	0.0001
Error	71	73.928685	1.041249		
Corrected Total	76	139.418831			

Source	DF	SS	Mean Square	F Value	Pr > F
LDOSE	1	51.646227	51.646227	49.60	0.0001
DRUG	2	13.755960	6.877980	6.61	0.0023
LDOSE*DRUG	2	0.087960	0.043980	0.04	0.9587

General Linear Models Procedure

Dependent Variable: RESPONSE

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	65.402186	21.800729	21.50	0.0001
Error	73	74.016645	1.013927		
Corrected Total	76	139.418831			

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
LDOSE	1	51.646227	51.646227	50.94	0.0001
DRUG	2	13.755960	6.877980	6.78	0.0020

Parameter		Estimate	T for HO: Parameter=0	Pr > $\|T\|$	Std Error of Estimate
INTERCEPT		-0.537010516	-1.30	0.1969	0.41235053
LDOSE		1.211112299	7.46	0.0001	0.16225105
DRUG	stand	0.238915716	0.80	0.4242	0.29725519
	formA	0.970022634	3.44	0.0010	0.28185554
	formB	0.000000000	.	.	.

SECTION F - STATISTICS FOR INDUSTRY AND QUALITY IMPROVEMENT

F1. A company manufactures gas burners for ovens. The specified length, from base to first gas outlet hole, for a burner is between 76.00 mm and 78.00 mm. The standard deviation of the length, estimated from past records when the plant has been perceived as working well, is 0.25 mm.

(i) Samples of 5 burners are taken at half-hour intervals, approximately. Set up a Shewhart chart for means, including 'warning' limits.

(ii) The first 10 means are:

sample number	1	2	3	4	5	6	7	8	9	10
mean	77.21	77.15	77.37	77.40	77.24	76.98	77.30	76.95	77.28	77.35

Plot these points on your chart. What action, if any, would you recommend taking if the process is difficult to set up and any stoppage is expensive?

(iii) Plot a CUSUM chart for the means in (ii). Explain how you would monitor this chart, sketching a V-mask and giving a brief practical justification of your choice of parameters for the mask.

(iv) After several weeks of satisfactory operation it is noted that the average within sample variance is 0.06mm^2, and the variance of the sample means is 0.02mm^2. Estimate the between sample variance and standard deviation. Comment briefly.

(v) What is *adaptive control*? When might it be appropriate? Would you recommend its adoption for this process?

Turn over

F2. An experiment was performed with the aim of improving the operational robustness of a card reader. The design factors were Amplifier Gain (AG), Current Limit (CL) and Spring Torque (ST) while Voltage Supply (VS) and Card Wear (CW) were the environmental factors. Each factor was included in the experiment at two levels: low (–) and high (+).

The measured response (y) was the amplified signal level from the magnetic head. The target value for y was 3.0. Eight designs were tested under the various environmental conditions and the results were as follows:

	Design factor settings			**Environmental factor settings**				
Design	**AG**	**CL**	**ST**	**VS:**	**–**	**+**	**–**	**+**	**MSD**
				CW:	**–**	**–**	**+**	**+**	
					Response (y)				
1	–	–	–		2.0	2.8	2.5	2.7	0.3450
2	+	–	–		1.5	3.2	3.2	1.5	1.1450
3	–	+	–		2.5	3.0	2.9	2.7	0.0875
4	+	+	–		2.1	3.4	3.3	1.7	0.6875
5	–	–	+		2.3	3.2	3.0	2.8	0.1425
6	+	–	+		1.8	2.5	3.3	1.5	1.0075
7	–	+	+		1.9	2.3	2.2	1.1	1.4875
8	+	+	+		0.6	2.8	3.2	0.7	2.7825

[Source: Industrial Statistics Research Unit, University of Newcastle upon Tyne.]

The mean squared deviation from target (MSD) for each design is given in the last column of the table. The objective was to minimise the MSD.

(i) Verify the calculation of the MSD in the first line of the table.

(ii) Sketch an interaction diagram for CL and ST, taking MSD as the outcome variable.

(iii) In order to assess the importance of the main effects and interactions on MSD, the appropriate contrasts were calculated. Confirm the contrast for the main effect of AG, and calculate the missing contrasts in the list below.

		Contrast Value
Main Effect	AG	0.89
	CL	0.60
	ST	0.79
Interaction	AG × CL	0.06
	AG × ST	0.19

(iv) Draw a normal score plot of these contrasts, and comment on your result. You may use the following expected values of standard normal order statistics for $n = 7$,

4th 0.000 5th 0.353 6th 0.757 7th 1.352

What advantage does a half-normal plot have over a normal score plot?

(v) Have you any reservations about your analysis?

F3. The yield of a chemical process depends on temperature and pressure. They can safely be adjusted within ranges 160° to 220° and $100kPa$ to $200kPa$ respectively. The current settings are 190° and $160kPa$. These were determined by the following experiment. The pressure was set at $150kPa$ and runs at different temperatures were made. A temperature of 190° resulted in the highest yield. The temperature was set at 190°, the pressure was varied, and $160kPa$ resulted in the highest yield.

(i) Explain to the plant manager, with the aid of a diagram, why this experiment may not have found the operating conditions that give the highest yield.

(ii) In a single replicate of a 2×2 factorial experiment, yield was regressed on temperature (x_1) and pressure (x_2), giving the following equation:

$$y = 44.50 + 0.09(x_1 - 190) + 0.21(x_2 - 160) \quad .$$

How many degrees of freedom are there for estimating the error? If x_1 is increased by 5° what change in x_2 corresponds to moving in the direction of steepest ascent? What is the estimated yield if these changes are made from the point (190, 160)?

(iii) Describe an efficient experimental design for two factors which would allow quadratic terms and interactions to be estimated.

(iv) The following regression equation resulted from the experiment described in (iii),

$$y = 48.81 - 0.89w_1 - 0.52w_2 - 1.53w_1^2 - 0.54w_1w_2 - 1.17w_2^2$$

where w_1 and w_2 are scaled deviations of temperature and pressure from a centre point. Find the values of w_1 and w_2 that give a maximum predicted yield. The residual mean square (s^2) was 1.85, and a 95% prediction interval for the yield from a single run at these values was quoted as:

$$\pm 2 \times \sqrt{1.85} \quad .$$

What is the basis of this quoted interval, and what criticisms do you have of it?

F4. (a) The cumulative distribution function of a Weibull distribution is,

$$F(x) = 1 - \mathrm{e}^{-(x/\beta)^{\alpha}} \quad , \quad \text{for } 0 < x \ .$$

Assume the following approximation concerning the expected value of the *i*-th order statistic in a sample of *n*,

$$F\left(E\left[X_{(i)}\right]\right) \approx \frac{i - 0.4}{n + 0.2} \ .$$

Explain how the parameters of the distribution can be estimated from a plot of the natural logarithms of the ordered sample values against $\ell n(-\ell n(1-(i-0.4)/(n+0.2)))$.

(b) Fourteen pieces of electric cable were subjected to an accelerated life test for 100 days. Three survived 100 days of testing, but the rest failed after

21 33 39 46 57 58 67 71 78 89 93

days respectively. Use graph paper to estimate the parameters of a Weibull distribution fitted to these data. Assuming the Weibull model, how many days would you expect the lifetimes of 99% of such test pieces to exceed?

Suppose that you wanted to estimate the Weibull parameters, for the above problem, using the maximum likelihood method. Explain how you proceed if the only method available for achieving the maximisation was one that did not depend upon knowledge of gradients. You are not expected to carry out this maximisation.